Psychometrika

A JOURNAL DEVOTED TO THE DEVELOPMENT OF PSYCHOLOGY AS A QUANTITATIVE RATIONAL SCIENCE

182

THE PSYCHOMETRIC SOCIETY · ORGANIZED IN 1935

VOLUME 9
NUMBER 3
SEPTEMBER
1944

PSYCHOMETRIKA, the official journal of the Psychometric Society, is devoted to the development of psychology as a quantitative rational science. Issued four times a year, on March 15, June 15, September 15, and December 15.

SEPTEMBER 1944, VOLUME 9, NUMBER 3

Offices of Publication: 23 West Colorado Avenue, **Colorado Springs, Colorado. Entered as** second class matter, September 17, 1940 **at the Post Office of Colorado Springs, Colorado,** under the act of March 3, 1879. **Editorial Office, 2539 Briarcliffe Avenue, Cincinnati,** Ohio.

Subscription Price:

To non-members, the subscription price is **$5.00 per volume of four issues.** Members of the Psychometric Society pay annual dues of **$5.00, of which $4.50** is in payment of a subscription to *Psychometrika.* **Student members of** the Psychometric Society pay annual dues of $3.00, **of which $2.70 is in payment for the journal.**

The subscription price to libraries and other institutions is **$10.00 per year; this price** includes one extra copy of each issue.

Applications for membership and student membership in **the Psychometric Society should** be sent to

THELMA GWINN THURSTONE
Chairman of the Membership Committee
University of Chicago
Chicago, Illinois

Membership dues and subscriptions are payable to

IRVING D. LORGE
Institute of Educational Research
Columbia University
New York, New York

(*Continued on the back inside cover page*)

Psychometrika

CONTENTS

VOLUME NINE SEPTEMBER 1944 NUMBER THREE

PSYCHOMETRIKA— VOL. 9, NO. 3
SEPTEMBER, 1944

THE DETERMINATION OF SUCCESSIVE PRINCIPAL COMPONENTS WITHOUT COMPUTATION OF TABLES OF RESIDUAL CORRELATION COEFFICIENTS*

LEDYARD R. TUCKER
PSYCHOMETRIC LABORATORY
THE UNIVERSITY OF CHICAGO

A procedure is presented for determining the successive principal components of a correlation matrix where it is not necessary to compute the successive tables of residual correlations. The original correlation matrix is bordered with a new row and column for each principal component that is determined.

The calculation of tables of residual coefficients of correlation has been one of the most laborious processes in the resolution of a set of variables into their principal components. Starting with an original matrix of correlations, R_1, the coefficients of the variables on the first principal component are determined. The entries in the table of residuals are computed by the formula

$$r_{2\cdot jk} = r_{1\cdot jk} - a_{j1}\, a_{k1}\,, \tag{1}$$

where $r_{2\cdot jk}$ is the residual coefficient, $r_{1\cdot jk}$ is the original correlation, a_{j1} and a_{k1} are the coefficients of variables j and k on the first principal component. The variable coefficients, a_{j2} and a_{k2}, on the second principal component are determined from the matrix R_2.

A simpler procedure for obtaining the second principal component is to border the original matrix R_1 by a new row and column as follows:

* This paper is one of a series of reports on the development of multiple factor analysis in the study of primary human abilities which have been supported by research grants from the Carnegie Corporation of New York and the facilities of the Psychometric Laboratory, which have been provided by the Social Science Research Committee of The University of Chicago.

$$\begin{array}{ccccccc|c}
r_{1\cdot 11} & r_{1\cdot 12} & \cdots & r_{1\cdot 1k} & \cdots & r_{1\cdot 1n} & & a_{11}\sqrt{k_1}i \\
r_{1\cdot 21} & r_{1\cdot 22} & \cdots & r_{1\cdot 2k} & \cdots & r_{1\cdot 2n} & & a_{21}\sqrt{k_1}i \\
\cdot & \cdot & \cdots & \cdot & \cdots & \cdot & & \cdot \\
\cdot & \cdot & \cdots & \cdot & \cdots & \cdot & & \cdot \\
\cdot & \cdot & \cdots & \cdot & \cdots & \cdot & & \cdot \\
r_{1\cdot j1} & r_{1\cdot j2} & \cdots & r_{1\cdot jk} & \cdots & r_{1\cdot jn} & & a_{j1}\sqrt{k_1}i \\
\cdot & \cdot & \cdots & \cdot & \cdots & \cdot & & \cdot \\
\cdot & \cdot & \cdots & \cdot & \cdots & \cdot & & \cdot \\
\cdot & \cdot & \cdots & \cdot & \cdots & \cdot & & \cdot \\
r_{1\cdot n1} & r_{1\cdot n2} & \cdots & r_{1\cdot nk} & \cdots & r_{1\cdot nn} & & a_{n1}\sqrt{k_1}i \\
\hline
a_{11}\sqrt{k_1}i & a_{21}\sqrt{k_1}i & \cdots & a_{k1}\sqrt{k_1}i & \cdots & a_{n1}\sqrt{k_1}i & & -k
\end{array}$$

R_{II}

where $k_1 = \sum_j a^2{}_{j1}$ and i is the imaginary number $\sqrt{-1}$. The first principal component of this enlarged matrix R_{II} is the second principal component of the original matrix R_1.

The foregoing relation can be demonstrated by considering the matrix A containing the coefficients on all of the principal components. This matrix has the properties that

$$AA' = R_1 \tag{2}$$

and

$$A'A = K, \tag{3}$$

where K is a diagonal matrix with the diagonal element for the mth principal component

$$k_m = \sum_j a^2{}_{jm}. \tag{4}$$

A new matrix, A_{II}, can be formed by adding a new row with a first entry of $\sqrt{k_1}i$ and all other entries of zero as follows:

$$\begin{array}{cccc}
a_{11} & a_{12} & a_{13} & \cdots \\
a_{21} & a_{22} & a_{23} & \cdots \\
\cdot & \cdot & \cdot & \cdots \\
\cdot & \cdot & \cdot & \cdots \\
\cdot & \cdot & \cdot & \cdots \\
a_{j1} & a_{j2} & a_{j3} & \cdots \\
\cdot & \cdot & \cdot & \cdots \\
\cdot & \cdot & \cdot & \cdots \\
\cdot & \cdot & \cdot & \cdots \\
a_{n1} & a_{n2} & a_{n3} & \cdots \\
\hline
\sqrt{k_1}i & 0 & 0 & \cdots
\end{array}$$

A_{II}

Then

$$A_{\text{II}} A'_{\text{II}} = R_{\text{II}} \tag{5}$$

and

$$A'_{\text{II}} A_{\text{II}} = K_{\text{II}} . \tag{6}$$

It will be noted that the only entry in K which is altered in going to K_{II} is the first diagonal. This first diagonal has a value of

$$k_{\text{II}\cdot 1} = \sum_j a^2_{j1} + (\sqrt{k_1}\, i)^2 , \tag{7}$$

which by (4) becomes

$$k_{\text{II}\cdot 1} = k_1 - k_1 = 0 . \tag{8}$$

Since K_{II} is a diagonal matrix, A_{II} contains the principal components of R_{II}, but equation (7) shows that the first column of A_{II} has a vanishing k and thus is not the first principal component of R_{II}. Hence the first principal component of R_{II} is the second column of A_{II}, which is the second principal component of R_1.

When a third principal component is desired, a matrix R_{III} can be formed by bordering R_{II} with a column and row with elements $a_{j2}\sqrt{k_2}\, i$. The diagonal element of this new row and column is $-k_2$. The proof that the first principal component of R_{III} is the third principal component of R is identical with the foregoing proof for R_{II}. It is desirable to know the variance of the correlations that is left unaccounted for after each factor, and this can be found by the relation

$$\sum_j \sum_k r^2_{m\cdot jk} = \sum_j \sum_k r^2_{1\cdot jk} - \sum_{p=1}^{m-1} k^2_p . \tag{9}$$

When Hotelling's iterative method (1) is being used and has been accelerated (2) by raising R to the power t so as to iterate on R^t, the same acceleration can be obtained by bordering R^t by a row and column with elements $a_{j1}\, k_1^{t/2}\, i$ and diagonal $-k^t$.

Table 1 represents a fictitious matrix R_1 which is used to illustrate the procedure for finding successive principal components without calculation of residual coefficients. The coefficients on the first principal component, a_{j1}, are also given in Table 1 as are values of k_1 and $\sqrt{k_1}$. R_{II} is given in Table 2. R_1 has been bordered by a row and column with side entries $a_{j1}\ \sqrt{k_1}\, i$. The diagonal entry of the new row and column is $-k_1$. Hotelling's iterative method was applied to R_{II} and Table 3 presents the first eight iterations. It will be noted that the successive trial values approach being proportional to the coefficients on the second principal component given in Table 2. The matrix R_{III} is given in Table 4. This matrix is found by bordering

R_{II} of Table 2 with a row and column with elements $a_{j2}\sqrt{k_2}\,i$ and diagonal of $-k_2$. The third principal component was determined from R_{III} by Hotelling's iterative method and is listed in Table 4.

REFERENCES

1. Hotelling, Harold. Analysis of a complex of statistical variables into principal components. *J. educ. Psychol.*, 1933, 24, 417-441, 498-520.
2. Hotelling, Harold. Simplified calculation of principal components. *Psychometrika*, 1936, 1, 27-35.

TABLE 1

R_{I}

	1	2	3	4	a_{j1}
1	.04	.04	.08	–.10	.0
2	.04	.29	.18	.08	.3
3	.08	.18	.56	.16	.6
4	–.10	.08	.16	.61	.6

$k_1 = .81 \qquad \sqrt{k_1} = .9$

TABLE 2

R_{II}

	1	2	3	4	I	a_{j2}
1	.04	.04	.08	–.10	.00*i*	.2
2	.04	.29	.18	.08	.27*i*	.2
3	.08	.18	.56	.16	.54*i*	.4
4	–.10	.08	.16	.61	.54*i*	–.5
I	.00*i*	.27*i*	.54*i*	.54*i*	–.81	.0

$k_2 = .49 \qquad \sqrt{k_2} = .7$

TABLE 3

Successive Iterations on R_{II}

	1	2	3	4	5	6	7	8
1	.000	.138	.382	.400	.400	.400	.399	.400
2	1.000	1.000	1.000	.667	.508	.445	.418	.406
3	.000	.621	.449	.667	.747	.780	.790	.796
4	.000	.276	–.949	–1.000	–1.000	–1.000	–1.000	–1.000
I	.000	.931*i*	.000	.000	.000	.002*i*	.000	–.022*i*

TABLE 4

R_{III}

	1	2	3	4	I	II	a_{j3}
1	.04	.04	.08	–.10	.00*i*	.14*i*	.0
2	.04	.29	.18	.08	.27*i*	.14*i*	.4
3	.08	.18	.56	.16	.54*i*	.28*i*	–.2
4	–.10	.08	.16	.61	.54*i*	–.35*i*	.0
I	.00*i*	.27*i*	.54*i*	.54*i*	–.81	.00	.0
II	.14*i*	.14*i*	.28*i*	–.35*i*	.00	–.49	.0

$k_3 = .20$

PSYCHOMETRIKA— VOL. 9, NO. 3
SEPTEMBER, 1944

FACTORING TEST SCORES AND IMPLICATIONS FOR THE METHOD OF AVERAGES

KARL J. HOLZINGER
THE UNIVERSITY OF CHICAGO

The general procedure and detailed steps for attaining complete factor analyses of scores are presented. Both orthogonal and oblique factors are considered. It is shown that a single average by conventional procedure gives an incomplete summarization of the data when the rank exceeds one. There should be as many averages as there are common factors.

I. *General Theory*

In factoring correlations it is customary to assume linear relationships between variables and factors of the form,

$$\begin{aligned} w_1 &= a_{11}\,G_1 + a_{12}\,G_2 + \cdots + a_{1m}\,G_m \\ w_2 &= a_{21}\,G_1 + a_{22}\,G_2 + \cdots + a_{2m}\,G_m \\ &\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot \\ w_n &= a_{n1}\,G_1 + a_{n2}\,G_2 + \cdots + a_{nm}\,G_m \end{aligned} \tag{1}$$

or

$$w_j = a_{j1}\,G_1 + a_{j2}\,G_2 + \cdots + a_{jm}\,G_m\,, \tag{1)'$$

where

$$w_j\,(j = 1\,,2\,,3\,,\cdots,n)$$

are the variables,

$$G_s\,(s = 1\,,2\,,3\,,\cdots,m)$$

are factors, and a_{js} are coefficients in these linear expressions. The variables and factors are usually taken in standard form (deviates from means divided by standard deviations). In the present discussion, however, all such variables will be taken in *normalized* form. Thus if X_j is a variable with mean M_j, then $x_j \equiv X_j - M_j$ and the normalized value becomes

$$w_j = \frac{x_j}{\sqrt{\sum x^2{}_j}}. \tag{2}$$

Variables in normalized form have the advantage of greater simplicity than standardized values, both analytically and geometrically. If $(i = 1, 2, 3, \cdots, N)$ indicates the range of individuals, then equation (2) may be written more precisely

$$w_{ji} = \frac{x_{ji}}{\sqrt{\sum_{i=1}^{N} x^2_{ji}}} \,. \tag{2$'$}$$

The correlation between two normalized variables w_j and w_k may then be written in the form

$$r_{jk} = \sum_{i=1}^{N} w_{ji}\, w_{ki} \,. \tag{3}$$

Obviously

$$\sum_{i=1}^{N} w^2_{ji} = 1 \tag{4}$$

and w_{ji} may be interpreted as direction cosines.

Inasmuch as a variable may be considered as a finite set of scores*, equations (1) may also be written as a set of relationships between scores on tests (or other variables) and factor scores. Equation (1)′ may then be written in the form

$$w_{ji} = a_{j1}\, g_{1i} + a_{j2}\, g_{2i} + \cdots + a_{jm}\, g_{mi} \,. \tag{5}$$

This expression will next be written in matrix form.

Let $$W_{ji} = \left\| \begin{matrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{21} & w_{22} & \cdots & w_{2N} \\ \cdot & \cdot & \cdot & \cdot \\ w_{n1} & w_{n2} & \cdots & w_{nN} \end{matrix} \right\|$$

denote the matrix of normalized scores.

* The reason the analysis of scores has been overlooked so long is probably due to the fact that alternate interpretations of the word "variable" have not been clear. If w_1 denotes the variable "height" we may imagine a continuum on which an indefinitely large number of values may be indicated. If a finite set of heights

$$W_{1i} : \| 62, 63, 69, 65, 64, \cdots, 68 \|$$

is given, this row matrix or "vector" may also be considered as a variable. It is this latter interpretation of "variable" that makes possible the geometric vector representation of variables, and suggests the factoring of scores instead of correlations.

Let
$$A_{js} = \left\| \begin{matrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{matrix} \right\|$$
denote the matrix of factor coefficients.

Let
$$G_{si} = \left\| \begin{matrix} g_{11} & g_{12} & \cdots & g_{1N} \\ g_{21} & g_{22} & \cdots & g_{2N} \\ \cdot & \cdot & \cdot & \cdot \\ g_{m1} & g_{m2} & \cdots & g_{mN} \end{matrix} \right\|$$
denote the matrix of factor scores.

Equation (5) may then be written as
$$W_{ji} = A_{js}\, G_{si} \tag{6}$$
or for purposes of calculation as
$$W_{ji} = A_{j1}\, G_{1i} + A_{j2}\, G_{2i} + \cdots + A_{jm}\, G_{mi}\,. \tag{6$'$}$$

There are two main problems in a factor analysis: the determination of A_{js} and of G_{si}. The calculation of A_{js} yields what is known as a "pattern" of the form (1) while the elements of G_{si} are the factor scores of the individuals. In factoring *correlations*, the matrix A_{js} is first determined by one of the several methods available, and then factor scores g_{si} are calculated directly or estimated. In factoring *scores* directly, this order is reversed, *i.e.*, the values g_{si} are computed by a simple process of averaging, and then the matrix A_{js} may be obtained.

II. *Orthogonal Factors*

Consider first the case of orthogonal centroid factors. The first centroid factor scores g_{1i} may be defined as those obtained by normalizing the totals (or averages) in the columns of W_{ji} (Table 3). This calculation yields the row matrix G_{1i}. The columns of the matrix A_{js} may then be found in the following manner: Postmultiply both sides of (6)$'$ by G_{i1} to obtain
$$W_{ji}\, G_{i1} = A_{j1}[G_{1i}\, G_{i1}] + A_{j2}[G_{2i}\, G_{i1}] \\ + \cdots + A_{jm}[G_{mi}\, G_{i1}]\,. \tag{7}$$
For orthogonal factors of the form here employed, the first square bracket becomes the identity matrix and the remaining terms vanish, yielding
$$W_{ji}\, G_{i1} = A_{j1} \tag{8}$$
as the first column of A_{js}.

The first term of the right-hand member of equation (6)′ may then be determined by the matrix product

$$A_{j1}\, G_{1i} = {}_1W_{ji}\,, \tag{9}$$

which may be interpreted as the part of W_{ji} attributable to the factor G_1.

The elements of ${}_1W_{ji}$ are next subtracted from those of W_{ji} to give the residual matrix $W_{ji} - {}_1W_{ji}$ of Table 4. Next the signs of the entries in certain rows of this last matrix are changed to "remove the centroid from the origin," and to make the new column totals as large as possible. The columns are then added and the totals normalized as before to yield the second-factor scores G_{2i} (Table 4).

Postmultiplying both sides of equation (6)′ by G_{i2} there results

$$W_{ji}\, G_{i2} = A_{j2}\,. \tag{10}$$

The product

$$A_{j2}\, G_{2i} = {}_2W_{ji} \tag{11}$$

gives the portion of W_{ji} attributable to the second centroid factor G_2. Subtracting the elements of ${}_2W_{ji}$ from those of $W_{ji} - {}_1W_{ji}$ yields the matrix of second residuals $W_{ji} - {}_1W_{ji} - {}_2W_{ji}$. This process may be continued until the final residuals are considered negligible. In the present illustration the original matrix is an artificial one of exactly rank two, so the second residuals of Table 5 are zero within errors of rounding.

The solution thus obtained will be identical with that by the usual centroid method based on correlations with unities in the diagonal. We shall not be concerned in the present paper with the vexing problems of communalities or "when to stop factoring," but merely present methods for factoring "whole scores" until the residual matrix contains entries considered practically unimportant.

III. *Steps in the Computation of the Centroid Solution*

The detailed steps in the calculations of the centroid solution will next be presented in such a manner that they may be followed by the student in routine fashion once he has mastered the simple operation of matrix multiplication.*

Step 1. Arrange the scores in a matrix with tests identified by rows and persons by columns (Table 1).

Step 2. Calculate the means of the rows of this matrix (Table 1).

Step 3. Subtract the mean of each row from each entry in the row of this matrix to obtain the deviates from the means. Obtain the

* Karl J. Holzinger and H. H. Harman, *Factor Analysis*, Appendix A. Chicago: University of Chicago Press, 1941.

sum of the squares of the entries in each row for subsequent calculation (Table 2).

Step 4. Divide each entry in Table 2 by the square root of the sum of the squares for each row to give the normalized deviates w_{ji} of Table 3. The sum of the squares of these values should be 1.000 for each row (check).

Step 5. Add the columns of W_{ji} (Table 3) to give the "column sums," and normalize these values in the manner employed to obtain the values w_{ji} in the body of Table 3. The last row of numbers in this table are the values g_{1i}, the first centroid factor scores. The sum of their squares should be checked as 1.000.

Step 6. Perform the matrix multiplication $W_{ji}\, G_{i1} = A_{j1}$ as illustrated below:

$$\left\| \begin{matrix} -.6708 & -.2236 & .2236 & .6708 \\ -.7071 & .4714 & .4714 & -.2357 \\ -.7487 & -.1248 & .2912 & .5823 \\ -.5607 & -.3271 & .1402 & .7476 \\ -.7720 & .4044 & .4779 & -.1103 \end{matrix} \right\| \times \left\| \begin{matrix} -.8312 \\ .0481 \\ .3855 \\ .3976 \\ \end{matrix} \right\| = \left\| \begin{matrix} .8997 \\ .6984 \\ .9601 \\ .8016 \\ .8015 \end{matrix} \right\|$$

The values a_{j1} at the right are the coefficients of the first centroid factor G_1 in the pattern (6).

Step 7. Perform the matrix multiplication $A_{j1}\, G_{1i} = {}_1W_{ji}$ as follows:

$$\left\| \begin{matrix} .8997 \\ .6984 \\ .9601 \\ .8016 \\ .8015 \end{matrix} \right\| \times \left\| \begin{matrix} -.8312 & .0481 & .3855 & .3976 \end{matrix} \right\| = \left\| \begin{matrix} -.7478 & .0433 & .3468 & .3577 \\ -.5805 & .0336 & .2692 & .2777 \\ -.7980 & .0462 & .3701 & .3817 \\ -.6663 & .0386 & .3090 & .3187 \\ -.6662 & .0386 & .3090 & .3187 \end{matrix} \right\|$$

The matrix on the right, ${}_1W_{ji}$, shows the part of the score matrix W_{ji} attributable to the first centroid factor.

Step 8. Subtract the elements of ${}_1W_{ji}$ from those of W_{ji} to obtain the residual matrix of Table 4; *e.g.* the top left element of Table 4 is $-.6708 - (-.7478) = .0770$, etc.

Step 9. The totals of the columns in Table 4 are zero, so the centroid must be removed from the origin. This is accomplished by changing the signs of the entries in the rows of the table to make the totals positive and large. In Table 4 the signs of rows 2 and 5 were changed because they had the largest negative values. A similar scheme might be followed with actual data.

Step 10. Calculate the sums of the columns and the normalized sums of the matrix of Table 4 (with sign changes) in same manner as in Step 5. The bottom row of Table 4 consists of the elements G_{2i}, the second centroid factor scores.

Step 11. Form the product $W_{ji}\, G_{i2} = A_{j2}$ as shown below:

$$\begin{Vmatrix} -.6708 & -.2236 & .2236 & .6708 \\ -.7071 & .4714 & .4714 & -.2357 \\ -.7487 & -.1248 & .2912 & .5823 \\ -.5607 & -.3271 & .1402 & .7476 \\ -.7720 & .4044 & .4779 & -.1103 \end{Vmatrix} \times \begin{Vmatrix} .1767 \\ -.6117 \\ -.2824 \\ .7175 \end{Vmatrix} = \begin{Vmatrix} .4364 \\ -.7155 \\ .2796 \\ .5978 \\ -.5979 \end{Vmatrix}$$

Step 12. Form the product $A_{j2}\, G_{2i} = {}_2W_{ji}$ as follows:

$$\begin{Vmatrix} .4364 \\ -.7155 \\ .2796 \\ .5978 \\ -.5979 \end{Vmatrix} \times \begin{Vmatrix} .1767 & -.6117 & -.2824 & .7175 \end{Vmatrix} = \begin{Vmatrix} .0771 & -.2669 & -.1232 & .3131 \\ -.1264 & .4377 & .2021 & -.5134 \\ .0494 & -.1710 & -.0790 & .2006 \\ .1056 & -.3657 & -.1688 & .4289 \\ -.1056 & .3657 & .1688 & -.4290 \end{Vmatrix}$$

This last matrix, ${}_2W_{ji}$, represents the part of the original scores W_{ji} attributable to the second centroid factor G_2.

Step 13. Subtract the matrix, ${}_2W_{ji}$, from the first residual matrix of Table 4 to obtain the matrix of Table 5 with both factors removed. With actual data, the above process may be continued until the residuals may be considered negligible.

Step 14. The two parts of the solution obtained are the factor pattern A_{js} calculated in Steps 6 and 11; and the factor scores of the individuals, G_{si}, obtained from the bottom rows of Tables 3 and 4. These two matrixes may be then written in the form of Tables 6 and 7. This last table illustrates one of the fundamental characteristics of factor analysis. The eight factor scores of Table 7 contain all the essential information about the four persons originally indicated in the matrix of Tables 1 or 3 by twenty scores.

IV. *Oblique Factors*

A solution involving correlated factors may be obtained much more simply than in the case of orthogonal factors in case the matrix W_{ji} can be sectioned (by rows) so that each sub-group of tests may be considered as measuring a single factor. The normalized totals for these sections then become the oblique factor scores denoted by l_{si} (Table 9). Such factors may be extracted all at once instead of one at a time as in the orthogonal case. Let

$$W_{ji} = A_{js}\, L_{si} \qquad (12)$$

represent the factor pattern for the oblique factors L_s. Postmultiplying both sides of (12) by L_{is} gives

$$W_{ji}\,L_{is}=A_{js}[L_{si}\,L_{is}]=A_{js}\,\phi_{ss}=\text{Structure*}=S_{js}\,, \qquad (13)$$

where ϕ_{ss} is the matrix of correlations between factors, and S_{js} is the matrix of the structure values, which are correlations between tests and factors. It will be noted that A_{js} is identical with S_{js} only in the case of orthogonal factors. In this case ϕ_{ss} equals the identity matrix I.

The pattern A_{js} can be obtained by postmultiplying both sides of equation (13) by ϕ_{ss}^{-1} to give

$$A_{js}=S_{js}\,\phi_{ss}^{-1}\,. \qquad (14)$$

The calculations are most readily done by the Doolittle† or similar method.

V. *Steps in the Calculation of Oblique Factors*

Step 1. Rearrange the rows of the normalized test scores W_{ji} of Table 3 according to the content of the tests or other criteria (Table 8). The basis for the grouping of the variables in this example was the agreement in algebraic signs of the scores. The points representing the variables in vector form thus fall into two distinct "sedeciments."‡

Step 2. Add the scores in these sections and normalize the totals as in Step 5 of the preceding section to give the oblique factor scores l_{1i} and l_{2i} (Tables 8 and 9).

Step 3. Postmultiply W_{ji} by the transpose of L_{si} to obtain S_{js} of equation (13) as shown below:

$$\left.\begin{array}{rrrr} -.7071 & .4714 & .4714 & -.2357 \\ -.7720 & .4044 & .4779 & -.1103 \\ -.6708 & -.2236 & .2236 & .6708 \\ -.7487 & -.1248 & .2912 & .5823 \\ -.5607 & -.3271 & .1402 & .7476 \end{array}\right\| \times \left\|\begin{array}{rr} -.7418 & -.6671 \\ .4392 & -.2276 \\ .4761 & .2207 \\ -.1735 & .6740 \\ & \end{array}\right\| = \left\|\begin{array}{rr} .9969 & .3096 \\ .9969 & .4541 \\ .3895 & .9998 \\ .5382 & .9846 \\ .2093 & .9833 \end{array}\right\|$$

Step 4. Obtain A_{js} from equation (14) by the Doolittle method, where

$$\phi_{ss}=L_{si}\,L_{is}=\left\|\begin{array}{rr} 1.000 & .383 \\ .383 & 1.000 \end{array}\right\|$$

The result of this computation is

[*] *Ibid.*, pp. 325-27.
[†] *Ibid.*, pp. 386-87.
[‡] *Ibid.*, p. 252.

$$A_{js} = \begin{Vmatrix} 1.0293 & -.0846 \\ .9645 & .0847 \\ .0077 & .9968 \\ .1888 & .9123 \\ -.1961 & 1.0584 \end{Vmatrix}$$

Step 5. From the values obtained in Steps 2 and 4 determine ${}_pW_{ji} = A_{js}\, L_{si}$ as "pattern scores":

$$\begin{Vmatrix} \mathbf{1.0293} & \mathbf{-.0846} \\ \mathbf{.9645} & \mathbf{.0847} \\ \mathbf{.0077} & \mathbf{.9968} \\ \mathbf{.1888} & \mathbf{.9123} \\ \mathbf{-.1961} & \mathbf{1.0584} \end{Vmatrix} \times \begin{Vmatrix} -.7418 & .4392 & .4761 & -.1735 \\ -.6671 & -.2276 & .2207 & .6740 \end{Vmatrix} = {}_pW_{ji}\,.$$

When this last multiplication for ${}_pW_{ji}$ is carried out, the entries are found to agree with those of Table 8 with a maximum discrepancy of .0003.

With actual scores the discrepancies might be large enough to

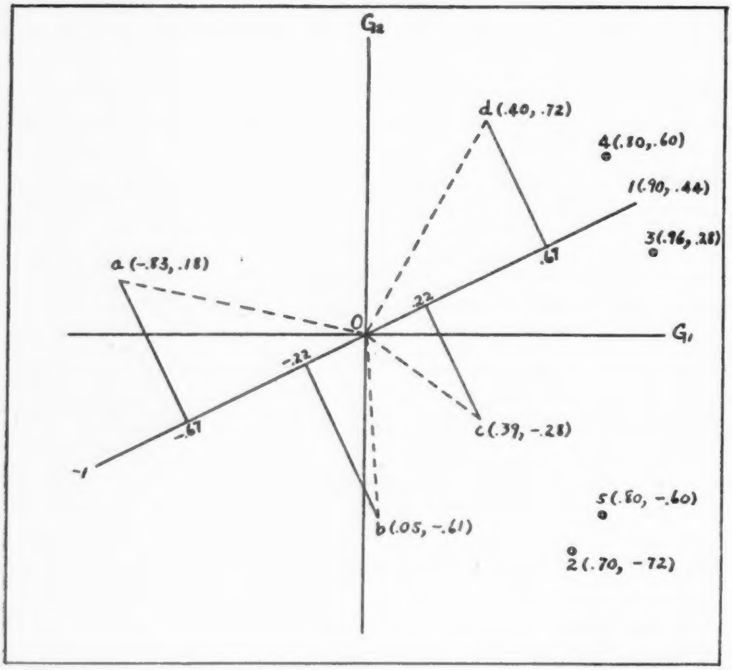

FIGURE 1

justify a resorting of the variables and recalculation of the whole solution.

VI. *Geometric Interpretation of the Data*

A simple geometric interpretation of the foregoing analysis may be made because all the data are contained in a two-space The normalized deviates may be considered as the coordinates of points equidistant from the origin. In Figure 1, the coordinates of the five test points with respect to the $G_1 - G_2$ axes are obtained from Table 6. The dotted lines to the points *a, b, c,* and *d* are the projections of the person axes on the plane here shown, the coordinates of the points being obtained from Table 7. The projections of these points in turn on the axis for Test 1 (for example) are also shown in the figure, the numerical values being those of the top row of Table 3. These values show the amount of Test 1 possessed by each of the four persons. The first coordinate of each of the points *a, b, c,* and *d* shows the amount of G_1 possessed by the person, while the second coordinate indicates the amount of G_2 he possesses.

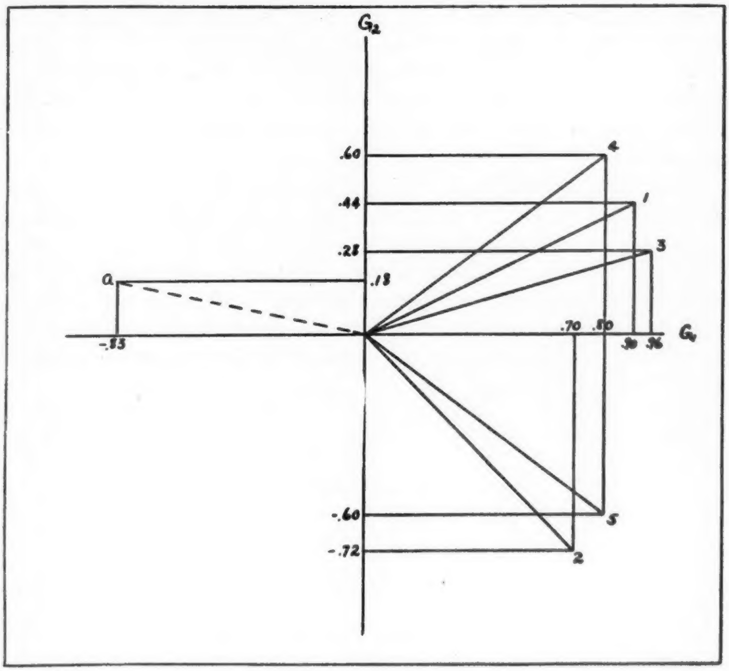

FIGURE 2

In Figure 2 the coordinates of the test points have been indicated as projections on the G_1 and G_2 axes. The projected axis Oa and the projections of its end point a on the G_1 and G_2 axes are also shown.

VII. *Some Implications of the Analyses*

These analytic and geometric interpretations illustrate short-comings of certain elementary statistical procedures. If the problem were to summarize the data of a table such as Table 3, the usual statistical procedure would be to add the columns and obtain averages. (In Table 3 the last row of numbers are merely normalized averages proportional to averages.) Here the ordinary investigation would stop, but it is quite apparent that only a *part* of the information about the data is thus obtained, *viz.,* G_{1i} . Geometrically, this means that although the data are in a two-space, only projections on the G_1 axis are considered. To complete the summarization of the data for the individuals, G_{2i} should be calculated, or geometrically, the projections on the G_2 axis should be made. A complete summarization could also be made, of course, by averaging the scores in sections as shown by the oblique analysis, but by either method, a complete analysis of the data must involve as many averages as there are factors.

The ordinary analyst might suppose that, even if several factors were involved in his data, a single average would somehow take them all into account. While this is true to some extent, much important information may still be ignored.

From these considerations it is apparent that a single average as a complete summarization is justified only if the data are of rank one; that is, only if one common factor is involved. To employ a single average for data of higher rank is to summarize the material incompletely. It would therefore appear that people who use the method of averages should be also familiar with the methods of factor analysis. Since nearly everybody uses the method of averages—perhaps that would be asking too much.

TABLE 1
Raw Scores of Four Persons on Five Tests

Test	Person				Mean
	a	*b*	*c*	*d*	
1	1	2	3	4	2.5
2	1	6	6	3	4.0
3	11	26	36	43	29.0
4	9	14	24	37	21.0
5	4	20	21	13	14.5

TABLE 2
Deviates from Means
x_{ji}

Test	Person				Sum of Squares
	a	*b*	*c*	*d*	
1	— 1.5	—0.5	0.5	1.5	5
2	— 3.0	2.0	2.0	— 1.0	18
3	—18.0	—3.0	7.0	14.0	578
4	—12.0	—7.0	3.0	16.0	458
5	—10.5	5.5	6.5	— 1.5	185

TABLE 3
Normalized Deviates
w_{ji}

Test	Person				Sum of Squares
	a	*b*	*c*	*d*	
1	— .6708	—.2236	.2236	.6708	.9999
2	— .7071	.4714	.4714	—.2357	1.0000
3	— .7487	—.1248	.2912	.5823	1.0000
4	— .5607	—.3271	.1402	.7476	.9999
5	— .7720	.4044	.4779	—.1103	1.0001
Column sums	—3.4593	.2003	1.6043	1.6547	17.3187
Normalized sums	— .8312	.0481	.3855	.3976	.9999

TABLE 4
First Residuals
$w_{ji} - {}_1w_{ji}$

Test	Person			
	a	*b*	*c*	*d*
1	.0770	—.2669	—.1232	.3131
2	—.1266	.4378	.2022	—.5134
3	.0493	—.1710	—.0789	.2006
4	.1056	—.3657	—.1688	.4289
5	—.1058	.3658	.1689	—.4290
Change signs of rows 2 and 5;				
Column sums	.4643	—1.6072	—.7420	1.8850
Normalized sums	.1767	— .6117	—.2824	.7175

TABLE 5
Second Residuals
$w_{ji} - {}_1w_{ji} - {}_2w_{ji}$

Test	Person			
	a	*b*	*c*	*d*
1	—.0001	.0000	.0000	.0000
2	—.0002	.0001	.0001	.0000
3	—.0001	.0000	.0001	.0000
4	.0000	.0000	.0000	.0000
5	—.0002	.0001	.0001	.0000

TABLE 6
Centroid Solution
A_{js}

Test	A_{j1}	A_{j2}
1	.8997	.4364
2	.6984	—.7155
3	.9601	.2796
4	.8016	.5978
5	.8015	—.5979

TABLE 7
Factor Scores for Two Centroids
g_{is}

Person	G_1	G_2
a	—.8312	.1767
b	.0481	—.6117
c	.3855	.2824
d	.3976	.7175

TABLE 8
Normalized Deviates
(Variables rearranged from Table 3)

Test	Person				Sum of Squares
	a	*b*	*c*	*d*	
2	—.7071	.4714	.4714	—.2357	
5	—.7720	.4044	.4779	—.1103	
Total	—1.4791	.8758	.9493	—.3460	3.975649
1	—.6708	—.2236	.2236	.6708	
3	—.7487	—.1248	.2912	.5823	
4	—.5607	—.3271	.1402	.7476	
Total	—1.9802	—.6755	.6550	2.0007	8.809318

TABLE 9
Values of l_{1i} and l_{2i}
(Normalized sub-totals of Table 8)

Factor	Person				Sum of Squares
	a	*b*	*c*	*d*	
L_1	—.7418	.4392	.4761	—.1735	.9999
L_2	—.6671	—.2276	.2207	.6740	.9998

PSYCHOMETRIKA— VOL. 9, NO. 3
SEPTEMBER, 1944

A NOTE ON CORRELATION CLUSTERS AND CLUSTER SEARCH METHODS

RAYMOND CATTELL
DUKE UNIVERSITY

Four methods of determining the clusters in a correlation matrix are described and compared. The choice of method has to be made according to the size of the matrix and the type of cluster sought. The relativity of clusters is emphasized and a distinction is drawn between phenomenal clusters and nuclear clusters. The relative utility of clusters and factors is briefly commented upon.

1. *The Setting of the Problem in Relation to Personality Research*

Since the study of individual differences aspired to become an exact science, psychology has had, as one of its major enterprises, the reduction of an almost endless variety of tests and ratings to a comparatively small number of representative variables. In this enterprise, especially as it concerned abilities, factor analysis has been unquestionably the main instrument. But since the wave of pioneer research entered the realm of personality variables there has appeared, among an appreciable number of workers (**2**, **6**, **7**, **9**, **10**, **11**, **12**, **14**, **15**) a preference for reduction into clusters rather than, or in addition to, factors. This may or may not prove justifiable, but since the discovery of clusters is, in any case, a valuable orientation-giving step, preliminary to factor analysis (especially in the new grouped centroid method of Thurstone, or Burt's analysis by sub-matrices (**1**) or in the present writer's use of a reduced "personality sphere" (**4**), the techniques of cluster study deserve more attention than has hitherto been given to them.

The present note offers some brief observations, arising from experience with a cluster analysis carried out on a larger scale than has previously been reported, concerning (1) the practical problems of establishing clusters, and (2) the relative utility of factors and clusters.*

* The research from which these observations primarily arise was directed to discovering the basic factors in 171 personality variables (**4**). This very comprehensive list of variables was first resolved into 35 clusters, of from three to ten variables in each, and which included all but two of the original 171 variables. The 35 cluster variables were then assessed in relation to a larger population than could be rated and handled with accuracy for the 171 variables, and the inter-correlations factored into some eleven factors (**5**).

2. *Four Methods of Determining Clusters*

When only one or two dozen variables are involved the procedure of discovering clusters by direct inspection of the correlation matrix is comparatively simple and adequate. But when the intercorrelations become more numerous, e.g., reaching 14,535 in our research, more systematic methods are necessary if some clusters are not to be overlooked. It is perhaps because of the smallness of the matrices so far encountered that so little has been written by way of practical guidance on systematic procedures.

From publications and from discussions with those recently engaged in cluster studies, one may gather that four principal methods are at present in use. They can contingently be labelled as follows:

(1) The Ramifying Linkage method.
(2) The Matrix Diagonal method.
(3) Correlation Profile Correlation (Tryon's method).
(4) The Approximate Delimitation (or Convergence) method.

All methods require, directly or indirectly, that the experimenter shall begin by fixing some lower limit to the magnitude of correlation coefficient which will be accepted as qualification for entry to a cluster. The qualifying variable must manifest either (a) a mean correlation with the other members of the cluster, and/or (b) its lowest correlation with any other cluster member, exceeding the given lower limit. In general (b) has proved more practicable and been more widely employed. Many studies have proceeded, for example, according to the rule that a variable can be added to a cluster only if it correlates 0.45 or more (uncorrected for attenuation) with every other variable therein. Typically this results in a *mean* inter-correlation within the cluster of between 0.5 and 0.7. When some minimum of this kind has been fixed, according to the general nature of the data and the range of coefficients found therein, cluster search becomes first a matter of looking for linkages, a linkage being defined as a "significant" correlation, i.e., something above the agreed minimum. One must begin, therefore, by marking on the correlation matrix, e.g., by colored encirclements, the coefficients which constitute linkages. From this point on the necessary steps differ, according to the method followed, as the accounts below indicate.

A. *The Ramifying Linkage method.* This is the simplest method, following first principles with pedestrian certainty. It has been used in most cluster studies, e.g., (7) and (12).

The first step is to make out for each variable a separate card, listing thereon, in order of their occurrence in the matrix, the other variables which have linkages with it. These lists we may call *Single*

Linkage lists. Thus one might find on the first card:

Variable A links with Variables *D, G, K, R, S, V, W.*

One then takes up the *D* Single Linkage list card and inspects it for linkages of *D* with the variables in the above list to the right of *D*, perhaps with the following result:

Variable D links with Variables *K, S, V, W.*

One next takes up the *K* Single Linkage list and searches it for relations with those to the right of *K*, perhaps with the following result:

Variable K links with *V* only.

A cluster of *A, D, K, V* has thus been established.

The whole procedure must now be repeated for Variable *B*, and so on through the list of variables, proceeding systematically in alphabetical order (or other extrinsic order of variables in the matrix) until all possibilities have been exhausted. The number of operations —the writing down of linkages on a card—in a matrix of, say, ten variables, is not great, but in a matrix of fair size, say, of seventy variables, it becomes enormous. If there are *N* variables in the matrix and the likelihood of one variable having a linkage with another is p (p usually being about one-tenth), one may show that the total number of operations, i.e., of inspection of coefficients and writing down of linkages, is approximately

$$N\left(\frac{N}{2}+\frac{Np}{2}\times\frac{Np^2}{2}\times\cdots\frac{Np^n}{2}\right) \text{ carried out until } \frac{Np^n}{2} \text{ equals unity.*}$$

This is true because each of the *N* variables requires on an average (1) $\frac{N}{2}$ coefficients to be inspected in making out the list; (2) $\frac{Np}{2}$ direct associates the lists of which have next to be inspected, and (3) $\frac{Np^2}{2}$ inspections arising from each of these, and so on until no more associates appear. Thus with 200 variables and a correlation level such that one r in ten is high enough to count as a linkage (i.e., $p = 1/10$), there are some 60,000 linkages to be inspected in abstracting all possible clusters from the table. Actually the above approximate formula may underestimate rather seriously because it assumes that the average frequency of linkages among themselves of the variables *D, G, K, R,* etc., linked to the first variable *A* is no greater than if they had no common relative A, but there is generally a tendency for variables related to a common variable to have greater than chance interrelatedness among themselves.

* This neglects the reduction by one variable in each operation and assumes an even frequency of linkage distribution.

B. *The Matrix Diagonal Method.* This method has been used by Burt (1), Cardall (2) in a thesis directed by Kelley, and others, first as a method of arranging tables for factor analysis and also as a method of discovering clusters *per se.** As before, the experimenter first marks on the correlation matrix those *r*'s high enough to count as linkages. (He may, further, grade them, according to two or three sizes of the coefficients, as in Diagram 1 below.) With or without making out single linkage lists for the variables, he then manipulates the order of the variables in the matrix in an attempt to *bring all the linkage correlations alongside the diagonal or as near to it as possible.* If the process can be successfully carried out, the resultant clustering is very clearly and strikingly recorded, as in the illustration of Diagram 1, taken from the findings of an actual research, based on the inter-correlations of a large number of interest tests.

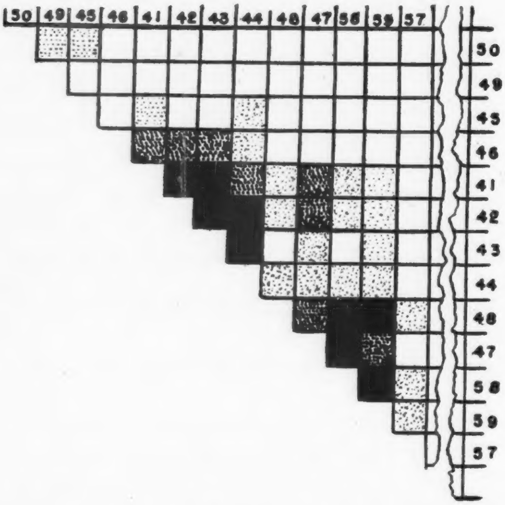

DIAGRAM 1

This is a corner fragment from a correlation matrix of 60 variables, each a measure of interest. The only correlations considered linkages have their cells shaded, the rest are blank. The order of variables has been re-arranged as indicated by the numbers at the edge.

r's above 0.60 are indicated by solid black.
r's from 0.50 to 0.59 are indicated by dark shading.
r's from 0.40 to 0.49 are indicated by light shading.

* It is also a necessary step in the search for types by inter-correlating persons instead of tests, in the well-known "Q-technique."

The drawbacks of this method are: first, that it cannot be reduced to any simple routine procedures; secondly, that it cannot be made "fool-proof" (indeed it may present a puzzle for a genius, becoming complex and difficult to the point of impossibility with 60 or more variables) ; and, thirdly, that it is not capable of dealing *at all* with the situation, by no means rare, in which two or three variables enter into three or more distinct clusters.

C. *Correlation Profile Correlation* (*Tryon's Method*). This method is too well known through Tryon's exposition (13) to need description here. Moreover, its theoretical implications cannot adequately be evaluated in so brief a discussion. A cluster by this definition is a set of variables which agree (beyond a certain arbitrary standard) in the profile (or rank order) of their correlations with all remaining variables of the matrix. This amounts to saying that the variables of a cluster must have their columns (or rows) of correlation coefficients in the matrix positively inter-correlating, above an agreed figure. Thus a set of variables chosen to satisfy Spearman's two-factor theory, and therefore falling into a "hierarchy," would constitute a cluster (at least, the higher members would.) Variables which showed a group factor in addition to the general factor would have more highly intercorrelated, i.e., more similar, profiles than would others, and would constitute a more distinguished cluster.

The relations of clusters established on this basis to those established on the simple inter-correlation basis employed by the great majority of researchers using clusters has never been sufficiently explored. Tryon's cluster might be called a "second-order cluster," for the variables are required to have similar profiles with regard to relations to other variables, instead of with regard to endowments of individuals in the variables concerned. There is no immediate reason to assume that the two concepts of cluster are identical, i.e., that the variables which behave in the same way to other variables will behave in the same way with respect to people. Indeed it is certainly possible to find instances in which the variables chosen for a first-order cluster would not be the same as those selected by the conditions of a second-order cluster. For example, in a well-formed Spearman hierarchy it is possible to select three or four variables at the bottom having low saturations with the general factor (and therefore low mutual inter-correlation) but having profiles of correlation with the remaining variables which have a similarity as great as or greater than that existing between variables high in the hierarchy and highly inter-correlated (especially if the latter have some small group factors breaking the hierarchy). However, in the instances of successful use of the method by C. M. Tryon (14) and in most examples tried out by the present writer, it seems true that a high degree of similarity of profile and high mutual inter-correlation go together. Assuming that this is a legitimate method of seeking ordinary clusters, we are, however, forced to conclude in the end that it does not contribute any short cut to cluster search methods. The matrix formed from inter-correlating variables by columns has as many coefficients as the original matrix. In a matrix of any size, therefore, it is still necessary to find some short process for detecting the clusters of mutually highly related variables.

D. *The Approximate Delimitation Method.* This is the title given, for lack of a briefer one, to the method invented for comparatively rapid, even if approximate, handling of cluster separation in large matrices. It was provoked by the problem of our own research (4) presenting a matrix with over 14,000 coefficients. Almost exactly 10% of the *r*'s in the matrix reached linkage level, so that, by the ramifying linkage method, some 60,000 *r*'s would have needed to be examined to determine the clusters, an almost prohibitive total.

The procedure is as follows:

(1) As in the ramifying linkage method a *Single Linkage List* is first prepared for each variable, showing the other variables with which it has *r*'s high enough to count as linkages, e.g., *A* links with *D, O, R, S, V, Z*; *B* with *C, D, L, N, W, Y, Z,* and so on.

(2) Each single linkage list, on a card, is systematically paired with each of the single linkage lists following it. That is to say, all possible comparisons, $\frac{N(N-1)}{2}$ in number, are made between the N single linkage lists. Whenever, in such a comparison, the two variables compared are found to have two or more common linkages, the two variables are listed on a new card. (Note that this has some resemblance to Tryon's method, for such variables have, as far as the rough test of linkage implies it, similar profiles). Thus *A* and *B* would begin a card because they share *D* and *Z*.

The new card is headed by the first variable of the pair, *A*, and accumulates under this heading all the variables which have two or more associates in common with the first. (Two proves a convenient minimum.) Of course, the associates which bind these variables together are not necessarily the same ones for different pairings of variables with *A*. If the variables on this new card list are also linked directly with each other, as evidenced by their being on the same *single linkage list,* they are underlined. (*B* above would not be underlined.) The new card lists are called the *Triangular Linkage Lists,* for each underlined variable thereon is certain to belong to at least a triad or "triangular linkage cluster" and almost certainly a tetrad, as Diagram 2 indicates.

An alternative procedure at this point is to examine, for the possession of common associates with the first variable, only those variables which fall in the *single linkage list* of the first variable. In this way one accumulates a *triangular linkage list* consisting only of the variables which would be underlined in the old triangular linkage lists. This offers a great saving, for perhaps only one-tenth of the paired comparisons mathematically possible among the variables will now need to be made. But at first we did not adopt this alternative, arguing that though the non-underlined variables of the longer triangular linkage lists fail to link (correlate highly enough) directly, they might eventually be found to have so many common associates, forming a cluster, that the failure of this single linkage could be overlooked. After all, the level of correlation accepted for entry into a cluster is arbitrary, and it is possible (and was indeed commonly found) that when one linkage only is absent an examination of the coefficient in the original matrix will show that it falls only negligibly below the prescribed level. It is a mistake, for this reason, to adopt any cluster search method that is not flexible, and some of the objections which might be raised against the present method on the grounds that it is not absolutely exhaustive should be considered in the light of this requirement.

However, later experience, as indicated below, showed that, at least in our material, no clusters were lost through working with the reduced triangular linkage lists, and since that procedure involves a great saving of time it may prove to be the better universal method.

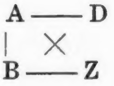

Diagram 2

TRIANGULAR LINKAGE LIST

If A and B are on the same list it follows (a) that they must be linked together, and (b) that both must be linked to at least two other variables such as D and Z, as indicated. Whether D is linked with Z is not known but such a linkage is probable.

(3) The triangular linkage lists are less numerous than the single lists; for not all variables are present even in clusters of three. The next step aims to bring together, as by a snowball adhesion of smaller fragments, whatever larger clusters can be formed from the triads. There is no single, clear, logically-defensible step by which

this can be done. Our procedure was to match in turn, systematically, the triangular linkage lists running into a single list those lists which had substantially (two-thirds) similar members. The new lists thus obtained we call the *Approximate Cluster Lists.* The process has reduced their number considerably (at least 50%) from the number of *triangular cluster lists.* They are very unequal in length. Further they can no longer be catalogued* under the heading of any single variable and its connections, for they are embryo clusters, or rather nebulae, out of each of which one or more clusters will condense when the final rigid criteria are applied to precipitate them.

(4) The last step consists in setting out the relations of variables in each approximate cluster list graphically in a square matrix. One need only refer to the single linkage lists to carry out this process, indicating each link by a cross in a cell, though one may, alternatively, decide to go back to the original matrix in order to fill in the cells with actual coefficients. Direct inspection will show the clusters present, though it may be necessary occasionally to rearrange the order of variables for greater clarity. One approximate cluster list commonly yields, by trimming in the matrix, a major cluster and some smaller ones. For example, a list of 17 variables yielded a cluster of 10, another of 6, three tetrads and eight triads. It is comparatively easy to lose the independent triads, but in our research we were interested in and finally recorded only clusters which were tetrads or larger (4).

Explained briefly and without much illustration, the above steps may appear somewhat complex, and their rationale somewhat obscure,

* The problem of cataloguing clusters and embryo clusters so that the belongingness of a variable, or the developing clusters themselves, can be readily traced, presents peculiar difficulties. Yet the whole organization of personality research, so far as it concerns collation, comparison and cross identification of clusters, either in traits or objective tests, would be facilitated a good deal by some efficient general system.

In the first place, at the triangular linkage stage, the embryo clusters are indexed under the first variable, which is scarcely more important than any other. At the approximate cluster list stage, the initial variable is quite unindicative of the character of the list and the experimenter has to depend on his memory of the general composition of each mass in order to identify or refer to it.

Even when the true cluster has finallv emerged to a stable life of its own, to a recognized pattern and perhaps "a local habitation and a name," a fairly difficult problem of indexing remains. Both embryo clusters and final clusters can be kept accessible only by having some rigid order, alphabetical or numerical, in the original variables and deriving the cluster index from this. If one then proceeds, in the exploration of linkages, systematically in one direction, the cluster aggregates can be indexed under the first variables in the cluster, and, next, under the second variable. Even so, one has difficulty in locating a cluster having a certain pattern of variables occurring beyond the second or third variable in its composition, so that eventually the only satisfactory indexing is one which lists opposite each variable, in a square table, all the clusters in which it participates.

but in actual practice they are entirely simple, provided one proceeds always through the variables very systematically and in one direction.

3. Choice of the Most Suitable Method

For the determination of operational clusters in the simplest fashion, as the above comments indicate, the choice lies between methods 1 and 4. The first of these, the ramifying linkage method, is entirely sure, but inexorably slow. The approximate delimitation method is excellent for detecting all clusters of appreciable size in a very large matrix, but it may let smaller clusters, of three or four variables, slip through its meshes. To make possible an appropriate choice between these methods it is necessary to compare them further.

The approximate delimitations method has the advantage of being to some extent adaptable in adjusting to time available and objectives required. For example, if one is interested only in finding the very large clusters, of, say, eight components or more, it may be necessary to list, in the second process of the method, i.e., the production of triangular linkage lists, only those variables which resemble each other by having at least *eight* common associated items (though it would, of course, be safer to fix a lower limit, at, say, six items).

Also, in most cases, it seems entirely safe to shorten process 2 by taking only the underlined linkages, as indicated above. For the linkage which is excluded from one triangular linkage list because it does not make a perfect triad is not entirely lost: it turns up from another angle in another list.

The approximate delimitations method operates by, as it were, roughly blocking in the cluster and its appendages, converging upon it from several directions, and leaving to the fourth process—that of drawing up a small matrix—the final chiselling out of the cluster from the block. The ramifying linkage method, on the other hand, picks up the cluster by one extremity and gradually disinters it by following up all the roots.

The latter is very sure, but gives every "phenomenal cluster" (see below) in detail, which is not always required. A more precise calculation of the comparative time and labor requirements of each will now be made and will show the very great gain to be obtained from using the former method.

There are, first, $\frac{N^2}{2}$ operations* in making the single linkage

* The same approximations will be made here as with the ramifying linkage method, namely, neglecting the reduction of combinations by one (i.e., $N-1$ is called N), and assuming an even distribution of frequency of linkages.

lists, as in the ramifying linkage method. Process 2 (making all possible paired comparisons of these to get triangular linkage lists) requires $\frac{N^2p}{2}$ operations, if only underlined variables are used. If, as seems usual, only two-thirds of the variables yield triangular linkage lists, the next step requires $\frac{(\frac{2}{3}N)^2}{2}$ operations, leading to the approximate cluster lists. The total is thus $\frac{N^2}{2} + \frac{N^2p}{2} + \frac{4N^2}{9}$. This exaggerates, because in the third step of condensing into approximate cluster lists not all the comparisons need to be made. Some lists coalesce relatively early, reducing the number of lists in the comparison operation. For 200 variables there are thus about 40,000 operations, as compared with 60,000 for the other method.

This comparison, however, does not do justice to the magnitude of the saving by the approximate delimitation method. In both methods a further procedure remains. The fourth process of the approximate delimitation method requires the making out and examining of about $\frac{N}{8}$ small matrices. The last process of the ramifying linkage method, on the other hand, requires an inter-comparison of all the clusters found; for, as the discussion below shows, the same individual clusters are picked up again and again from different angles. This process of simplification (requiring a fairly elaborate and vigilant bookkeeping) is a more difficult and prolonged one than the fourth process of the approximate linkage method.

4. The Nature of Clusters: Nuclear and Phenomenal Clusters

Any final evaluation of the utility of different search methods depends upon the type of cluster one is seeking. The discussion of methods is therefore appropriately brought to a close by considering the second topic of this paper, namely, the definition of the varieties of correlation clusters.

Current theoretical discussions on clusters, in personality or cognitive performances, generally proceed on the assumption that clusters are well defined, discrete entities, and of one kind only. Any actual exploration of correlation matrices, however, in which correlations above a certain size are represented by linkages, reveals almost invariably a somewhat bewildering network of partly linked and more or less overlapping clusters. Fortunately, in most data, the overlapping is not purely haphazard or evenly distributed: there is rather a tendency for it to occur about a more limited number of nuclei. The whole discussion of this subject can therefore be clarified if we distinguish two kinds of clusters, which, for lack of better terms, we may call *nuclear* and *phenomenal* clusters, and which may be illustrated in the diagram below.

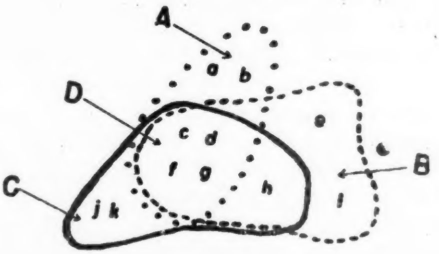

DIAGRAM 3

A First Phenomenal Cluster, containing variables *a, b, c, d, f, g*.
B Second Phenomenal Cluster, containing variables *c, d, e, f, g, h, i*.
C Third Phenomenal Cluster, containing variables *c, d, f, g, h, j, k*.
D Nuclear Cluster, containing variables *c, d, f, g*.

This represents three phenomenal clusters, A, B, and C, within each of which all the variables have mutual correlations of a high level. They overlap in the variables *c, d, f, g*, but no other variables are common to all three. Thus *e* correlates sufficiently with *c, d, f,* and *g*, but not with *a* and *b* or with *j* and *k*. The cluster *c, d, f, g* may therefore be called a core or nucleus, and the outlying portions of the phenomenal clusters might be called the "appendages" of this nuclear cluster.

The ramifying linkage method gives all three of the above phenomenal clusters separately, at widely separated points in the systematic search, requiring the experimenter to recognize the kinship or overlap and to superimpose the clusters in order to discover the nuclear cluster. The approximate delimitation method, on the other hand, gives a single matrix, covering all the variables—*a* through *k*. Within this matrix the relationships can be at once seen, and from it, paring like a gem-cutter upon the rough stone, the experimenter can isolate the individual clusters.

If one is aiming at nuclear clusters only, the exhaustive ramifying linkage method is a waste of time. Yet if one attempts to shorten it by not pursuing the method exhaustively, e.g., by following up only the first half of the leads, the results are misleading. Undue importance in determining the ultimate shape of the cluster is given to the variables which, by the accident of alphabetical order, happen to be first on the scene. Variables which fail to correlate to the specified amount with these first variables are dropped, with all their ramifications unexplored, though they might in fact correlate satisfactorily with all later variables in the cluster.

Whether the true goal of research is to find phenomenal clusters or nuclear clusters is a question to be decided by the problem and by circumstantial considerations which lie outside the scope of this note. For some personality problems it may be important to know the phenomenal clusters—to know, for example, that though variables *a* and *b* on the one hand and *j* and *k* on the other adhere as "appendages" to the same temperamental *nuclear cluster c, d, f, g,* they are in fact alternative derivatives, not appearing simultaneously in the same personality. (They are probably opposite aspects of what Burt (1) calls a "bipolar factor.") At the present stage of personality research, however, the mapping of appendages and conditional fringes should perhaps be considered secondary to the urgently necessary task of establishing the main nuclear clusters. Moreover, until these are confirmed by several researches and delimited with some accuracy, the appendage outlines and the variables on the fringes ought scarcely to be involved in serious hypotheses, for they may represent nothing more than the effects of sampling errors pushing correlations now above, now below the accepted level for linkage (among variables on the borderline of admission to a cluster). Finally, in deciding between these two methods and objectives, one must bear in mind that if the object of research is to reduce the number of variables with which further research has to deal, the ramifying linkage method, with its extensive harvest of phenomenal clusters, fails. Indeed in a majority of the researches we have studied the phenomenal clusters are actually somewhat more numerous than the variables from which they are isolated.

Further consideration of varieties of clusters will show that there are not only clear phenomenal and nuclear clusters, but various degrees of nuclearity, requiring, perhaps, special terms. For the purpose of our own research (4) we listed only strictly nuclear clusters (*c, d, f,* and *g* in the example above), expressing the chief phenomenal clusters as "alternative appendages" (e.g., *h, h—e—i, a—b,* and *j—k*), always recorded alongside the nuclear cluster; but in other researches some record of intermediate nuclear clusters, e.g., *c—d—f—g —h,* might be appropriate.

All questions concerning (a) the indexing of clusters as dependent or independent, (b) the decision as to degrees of nuclear belongingness, and (c) agreement on the number of clusters to which the matrix can be reduced, are very intimately tied up with the level of correlation accepted as the criterion for admission to a cluster. The situation is really no different from that facing the cartographer, when he has to decide to what land masses the terms hill and mountain are to be attached. A matrix which yields, in a certain area, two distinct clusters when a high level of correlation is demanded may yield only one large cluster when the level is lowered, as two islands become one when the tidal level falls. (See Diagram 4 below, in which this relativity of clusters is rendered particularly obvious through the spatial expression of correlations as cosines.)

5. *Clusters or Factors?*

It seems to be the contention of those recent researches which have preferred cluster analysis to factor analysis that while clusters reduce the number of variables practically as effectively as factors, they enjoy a greater reality than factors. We shall debate this; but before doing so we shall admit one very real advantage to clusters, namely, that they permit the results of different researches to be relatively easily combined. When the results of several different cluster researches, on the same variables and similar populations, are collated, it is possible to see at once where the results of one research are approximately confirmed by those of another, for no problem of rotation or of factor system arises. And when partly different sets of variables are used in successive researches it is possible to augment the description of clusters arrived at in the first research by adding those variables in the second which belong to recognizably similar clusters. (Those accumulated clusters will need later confirmation, and proof that no variables stray outside the accepted angle.) Since few researches use exactly the same sets of variables, this possibility of cumulative knowledge through overlapping researches is very attractive. With factor analysis, on the other hand, where the factorization varies with the mathematical system adopted and with the trait population, and where experimenters may not use the same rotation and orientation of axes, the utilization in a single integrated conclusion of results from different researches (without going back to the original correlations) is often practically impossible.

But the notion that the cluster enjoys an *absolute* reality and stability cannot be left undisputed. In the first place, the membership and shape of a cluster is affected as much as, and commonly more than, a factor, by sampling errors of population and by varying reliabilities of the tests. Secondly, even when freed of such errors, clusters have intrinsic indeterminacy. Consider the situation in Diagram 3 below, in which correlations are represented geometrically (for simplicity in two dimensions). In a research using variables *a* through *k* and a cluster criterion of a minimum correlation of 0.75 there are three clusters, A, B, and C. If the criterion is lowered to 0.40 there are two clusters, A′ and B′, overlapping. If, with the first criterion, the variables are increased by adding *l* and *m*, correlating in the manner shown, there appears a series of continuously overlapping clusters from *a* to *k*, such that only by some further arbitrary step is it possible to determine what clusters shall be said to exist. In some situations, incidentally the arbitrariness of cluster demarcation may be avoided by borrowing from factor analysis Holzinger's B-coefficient technique, so that a cluster is defined by a density of grouping of the variable lines (in Diagram 3) relative to the general density in that region (8). This method could, in fact, be substituted anywhere for the rigid, minimum-correlation criterion adhered to in this article, so that clusters might have widely different degrees of internal inter-correlation, depending on the surrounding density in their field. But it would not solve the present problem. In short, a cluster may become quite as arbitrary and quite as indeterminable by the nature of the given correlations alone, as any factor. Fortunately clusters are not so likely to run together in the above confusing fashion in *n*-dimensional space as in a two-dimensional diagram, and with a fairly high criterion of cluster belongingness, e.g., a minimum *r* of 0.6, problems of separating clusters seldom arise.

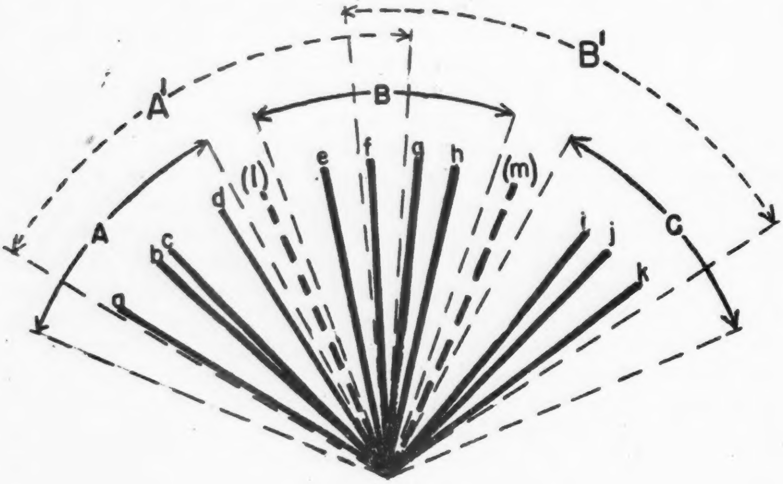

Diagram 4

Elsewhere (3) the present writer has put forward reasons for believing that a psychological functional unity, i.e., an entity real in some psychological causal sense, is more likely to be represented as a factor than as a cluster in observed covariation relationships. A cluster, it is true, may arise from a factor, in the sense of being constituted by those traits which are highly loaded by that factor; but a cluster is at least as likely to represent instead a set of variables which share the cumulative overlap of several factors, not one of which is highly represented in them. If the regions(sets of variables) of overlap of factors are affected by local and circumstantial conditions whereas the factors represent real influences, it follows that many clusters will be transient, unstable phenomena, while factors will be more constant in appearance and in meaning. For example, general intelligence may be a factor and the pattern of a classical education may be a factor, but the cluster which appears as a high correlation of intelligence items and knowledge of classics items, and which guided many scholastic and civil service appointments in the last century, may disappear with the disappearance of the practice of directing the most intelligent students into classical studies.

The maintenance of cluster analysis and factor analysis in their true roles and relationships may perhaps be best assured by the brief dictum that clusters are essentially representations at the descriptive level, and as such are little better than straight statements of the correlation coefficients, whereas factors are statements at the interpretive level. *If* the interpretations are correct the factors have more permanent value and far wider utility. But until there is more general agreement on the rotation of axes it may be desirable to publish analyses of correlation matrices both in cluster analysis and in factor analysis form; for the former will preserve results in a shape suitable for immediate collation with those of other researches. Secondly, as in the research from which these observations arise, (4, 5) the cluster analysis may be used as a first reduction of variables, to provide a briefer list upon which a factor analysis can be more practicably carried out than on the original plenary list. Naturally this factor analysis at one remove will not yield all the factors required to account for the variance of all variables, of the original plenary list, but it will provide the more important ones for the major structuring of the field, i.e., a broad framework within which the findings of later, more restricted and local factor analyses can be fitted in perspective.

References

1. Burt, C. L. The factors of the mind. London: Univ. London Press, 1940.
2. Cardall, A. J. A test for primary business interests. D. Educ. Thesis. Dept. Educ. Library, Harvard Univ., 1941.
3. Cattell, R. B. The description of personality. I. Foundations of trait measurement. *Psychol. Rev.*, 50, 559-594, 1943.
4. ———. The description of personality. II. Basic traits resolved into clusters. *J. abn. soc. Psychol.*, 38, 476-506, 1943.
5. ———. The description of personality. III. Principles and findings in a factor analysis of the personality sphere. (Publication to be announced).
6. Flemming, E. G. A factor analysis of the personality of high school leaders. *J. appl. Psychol.*, 5, 596-605, 1935.
7. Horn, D. A study of some syndromes of personality. *Charact. and Person.*, 12, June, 1944.
8. Holzinger, K. J., and Harman, H. H. Factor analysis. Chicago: Univ. Chicago Press, 1941.
9. Maslow, A. H. A test for dominance feelings in college women. *J. soc. Psychol.*, 12, 255-270, 1940.
10. Maurer, K. M. Patterns of behavior of young children as revealed by a factor analysis of trait "clusters." *J. genet. Psychol.*, 59, 177-188, 1941.
11. McCloy, C. H. A factor analysis of personality traits to underlie character education. *J. educ. Psychol.*, 27, 375-384, 1936.
12. Sanford, R. N. Personality patterns in school children. In Barker, R. G., Kounin, A. S., and Wright, H. F., Child behavior and development. New York: McGraw-Hill, 1942, pp. 567-589.
13. Tryon, R. C. Cluster analysis. Ann Arbor: Edwards, 1939.
14. Tryon, C. M. Evaluations of adolescent personality by adolescents. In Barker, R. G., Kounin, A. S., and Wright, H. F., Child behavior and development. New York: McGraw-Hill, 1942.
15. Williams, H. M. A factor analysis of Berne's "Social behavior patterns in young children." *J. exp. Educ.*, 4, 142-146, 1935.

PSYCHOMETRIKA— VOL. 9, NO. 3
SEPTEMBER, 1944

FUNDAMENTAL FACTORS OF COMPREHENSION IN READING

FREDERICK B. DAVIS
COOPERATIVE TEST SERVICE OF THE AMERICAN COUNCIL ON EDUCATION*

A survey of the literature was made to determine the skills involved in reading comprehension that are deemed most important by authorities. Multiple-choice test items were constructed to measure each of nine skills thus identified as basic. The intercorrelations of the nine skill scores were factored, each skill being weighted in the initial matrix roughly in proportion to its importance in reading comprehension, as judged by authorities. The principal components were rather readily interpretable in terms of the initial variables. Individual scores in components I and II are sufficiently reliable to warrant their use for practical purposes, and useful measures of other components could be provided by constructing the required number of additional items. The results also indicate need for workbooks to aid in improving students' use of basic reading skills. The study provides more detailed information regarding the skills measured by the *Cooperative Reading Comprehension Tests* than has heretofore been provided regarding the skills actually measured by any other widely used reading test. Statistical techniques for estimating the reliability coefficients of individual scores in principal-axes components, for determining whether component variances are greater than would be yielded by chance, and for calculating the significance of the differences between successive component variances are illustrated.

The application of techniques of factorial analysis to the investigation of reading has been attempted several times. Feder (11), Gans (12), and Langsam (23) have published studies that employed Thurstone's centroid method, and unpublished studies have been made by Bedell and Pankaskie. So far as the writer is aware, the study reported here is the first to make use of tests especially constructed to measure the mental skills in reading comprehension that are considered of greatest importance by authorities in the field.**

The most important step in a study that employs factorial procedures for the investigation of reading comprehension is the selection of the tests the scores of which are to be factored. Unless these tests provide reasonably valid measures of the most important mental skills that have to be performed during the process of reading, the application of the most rigorous statistical procedures can not yield meaningful and significant results. The importance of this point can hardly be overstated.

* On leave for military service.

** For a detailed presentation of the basic data of this study, see (8).

As the first step in the present study, a careful survey was made of the literature to identify the comprehension skills that are deemed most important by authorities in the field of reading. A list of several hundred specific skills was compiled, many of them overlapping. This list of skills was studied intensively by the writer in order to group together those that seemed to require the exercise of the same, or closely related, mental skills. The objective was to obtain several groups of skills, each one of which would constitute a cluster having relatively high intercorrelations and relatively low correlations with other clusters of skills.

Nine groups of skills were sorted out and labeled. For the purposes of this study, they are regarded as the nine skills basic to comprehension in reading. Included within them is the multitude of specific skills considered important by the authorities consulted. These nine basic skills are as follows:

1 Knowledge of word meanings
2 Ability to select the appropriate meaning for a word or phrase in the light of its particular contextual setting
3 Ability to follow the organization of a passage and to identify antecedents and references in it
4 Ability to select the main thought of a passage
5 Ability to answer questions that are specifically answered in a passage
6 Ability to answer questions that are answered in a passage but not in the words in which the question is asked
7 Ability to draw inferences from a passage about its contents
8 Ability to recognize the literary devices used in a passage and to determine its tone and mood
9 Ability to determine a writer's purpose, intent, and point of view, i.e., to draw inferences about a writer

To provide a measure of each one of these nine basic skills, a large number of five-choice objective test items were constructed. All possible care was taken to obtain items that measured only one rather than several of the nine skills. However, it was recognized that skill 1 (knowledge of word meanings) is basic to the measurement of all the other skills, since to read at all one has to recognize words and understand their meanings, and that some overlapping of skills 2-9 is inevitable.

Since the final forms of the reading-comprehension tests used in this study were to be the published forms of Tests C1 and C2 of Form Q of the *Cooperative Reading Comprehension Tests,* practical considerations [notably the requirements of the procedure for obtaining three equivalent "scales" in the tests (6)] determined in some measure the number of items testing each basic skill that could be used. An effort was made, however, to include the proportion of items testing each one of skills 2-9 that conformed to the judgments of authorities in the field of reading.

To obtain the intercorrelations of scores in the nine basic reading skills selected for measurement, 240 multiple-choice items were administered to a large number of freshmen in several teachers colleges.* The students were told to mark every item and were allowed an unlimited amount of time. By this means, the influence of speed of reading was removed and the effects of mechanical difficulties in word perception were minimized. Of the 541 students tested, 421 actually answered every item, and, when proof was obtained that this group constituted a representative sample of the entire 541 students tested, the scores of only these 421 pupils were used in the factorial analysis. In addition to the intercorrelations of the scores, the correlations between sex and scores in each of the nine skills were computed. As would have been expected, the correlations with sex were all insignificantly different from zero. This being so, there was no need to partial out the influence of sex before making a factorial analysis.

Table 1 shows the intercorrelations of the scores in the nine basic reading skills, and their relationships with sex.

TABLE 1

Intercorrelations, Means, and Standard Deviations of Raw Scores in the Nine Basic Reading Skills, and Their Relationships with Sex
($N = 421$)

Skill	1	2	3	4	5	6	7	8	9	Sex*	Mean	σ
1		.72	.41	.28	.52	.71	.68	.51	.68	.03	23.77	**11.61**
2			.34	.36	.53	.71	.68	.52	.68	–.07	12.70	**3.25**
3				.16	.34	.43	.42	.28	.41	–.01	4.20	**1.73**
4					.30	.36	.35	.29	.36	–.03	2.97	**1.10**
5						.64	.55	.45	.55	–.04	18.10	**2.46**
6							.76	.57	.76	–.01	25.67	**5.67**
7								.59	.68	.06	28.46	**5.81**
8									.58	–.05	6.75	**1.86**
9										–.05	15.19	**4.07**

* A pcsitive coefficient in this column indicates that the men obtained a higher mean score than the women.

* Every freshman in all of the teachers colleges of the State of Connecticut and every freshman in two of the Massachusetts State Teachers Colleges comprised the sample tested. The testing was done about a month after the beginning of the school year.

The intercorrelations of the nine basic skills range from .16 to .76, the values reflecting in part their true relationships and in part the differences in their reliability. The reliability coefficients of the scores in the nine skills are shown in Table 2.

TABLE 2

Reliability Coefficients of Raw Scores in Each of the Nine Basic Reading Skills*

Skill	r_{1I}	N	Number of Items
1	.90	100	60
2	.56	100	20
3	.44	100	9
4	.18	421	5
5	.55	100	22
6	.77	100	42
7	.63	100	43
8	.64	100	10
9	.71	100	27

* The division of each test into two halves was accomplished in this case by arranging the items in order of difficulty and assigning alternate items to each half. It will be recalled that speed had no influence on these scores. The reliability coefficient for skill 4 is based on 421 cases; the reliability coefficients for the other skills are based on a representative sample of 100 cases drawn from the 421 available.

As would be expected in view of the widely different lengths of the tests used to measure the nine basic reading skills, their reliability coefficients differ considerably. For even the least reliable, however, the reliability coefficient is substantially and significantly greater than zero.

Subjective judgment had forecast relatively high correlations between skill 1 and each of skills 2-9. Inspection of Table 1 in the light of the data in Table 2 reveals this to be so. It is apparent that skill 1 constitutes the largest element common to all of the other initial variables; hence, it may be of interest to study the intercorrelations of skills 2-9 when skill 1 is held constant. These partial coefficients are shown in Table 3.

TABLE 3

Partial Correlation Coefficients Among Skills 2-9, Skill 1 Being Held Constant ($N = 421$)

Skill	3	4	5	6	7	8	9
2	.09	.23	.26	.40	.38	.26	.37
3		.05	.16	.22	.22	.09	.20
4			.19	.23	.22	.17	.24
5				.45	.32	.26	.32
6					.53	.33	.53
7						.38	.40
8							.38

Perhaps the most surprising feature of the data in Table 3 is the small size of the coefficients. After making due allowance for the attenuation resulting from the comparatively low reliability coefficients of some of the variables, it is apparent that reading comprehension, as measured by the nine basic reading skills, is not a unitary ability. From the correlations it appears probable that a mental ability present to the greatest extent in skills 6, 7, and 9 is second most important in producing the intercorrelations shown in Table 1. To explore this matter, a factorial analysis was undertaken, using the method described by T. L. Kelley (22).*

The initial matrix of variances and covariances used in the factorial analysis is presented in Table 4.

TABLE 4

Initial Matrix of Variances and Covariances*

Variable	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
x_1	134.70	27.01	8.16	3.65	14.77	46.88	45.78	11.04	32.07
x_2		10.56	1.94	1.29	4.22	13.03	12.90	3.17	8.93
x_3			3.01	0.31	1.44	4.24	4.24	0.90	2.91
x_4				1.22	0.82	2.25	2.25	0.59	1.63
x_5					6.05	8.93	7.85	2.07	5.53
x_6						32.17	24.89	5.96	17.42
x_7							33.75	6.33	16.00
x_8								3.46	4.42
x_9									16.54

* Variances are shown in the diagonal cells. The Kelley method would be equally applicable if the scores in variables 1-9 were transformed into standard measures. In this case, the variance in each diagonal cell would be 1 and the covariances would be identical with the intercorrelations shown in Table 1. The resulting matrix would undoubtedly present a more familiar appearance to many students. Each one of the basic reading skills would then have been weighted equally for purposes of factorial analysis. However, authorities in the field of reading quite reasonably do not judge each one of the basic skills to be of equal importance in the process of reading comprehension. Of the many possible factorial analyses (using different weights), that analysis which appears to have unique merit is a principal-axes solution based on a matrix of variances and covariances in which the initial test variances are weighted to correspond with their relative importance in the process of reading, as determined by the pooled judgment of authorities. That is the type of factorial analysis that it was intended should be performed in the present study, but practical considerations resulted in some modifications in the relative weights of the nine initial variables.

For purposes of comparison, the Kelley method was used to perform a factorial analysis of the correlation matrix shown in Table 1 (excluding sex) with unit variances in the diagonals. A comparison of the factor loadings derived from the two principal-axes analyses and from a centroid analysis of the same data is now in preparation.

In Table 5 are presented the coefficients of each of the initial variables (the nine basic reading skills) that yield the nine independent components obtained by factorial analysis. The design shown in Table 5 is one of the most interesting that has been obtained by factorial techniques.

* For this study it was desirable to obtain the factor loadings of all significant components rather than the loadings for only the two or three largest components; hence a fairly large number of subjects was tested and Kelley's method was selected as being most suitable for use.

TABLE 5

Coefficients of Each of the Initial Variables That Yield Scores in the Nine Independent Components

(Factor Loadings of Skills 1-9 in Components I-IX)

Components	I	II	III	IV	V	VI	VII	VIII	IX	
Variance	192.270	22.824	8.657	5.282	3.828	3.306	2.327	1.956	1.006	
Skills										Variance
1	.813	–.571	–.064	–.033	–.082	.006	–.016	.001	.011	134.699
2	.184	.124	–.005	–.003	.971	–.019	–.017	–.028	–.076	10.563
3	.057	.054	–.001	.000	–.000	.000	.997	.000	–.004	3.009
4	.027	.048	–.000	.000	.067	.000	.000	.000	.996	1.220
5	.107	.149	.152	–.003	–.022	.970	–.014	–.024	–.012	6.050
6	.341	.469	.567	–.531	–.129	–.204	–.044	–.001	–.023	32.169
7	.336	.580	–.719	.008	–.147	–.020	–.051	–.091	–.028	33.752
8	.078	.105	–.001	.141	–.000	.000	–.010	.981	–.007	3.456
9	.233	.253	.366	.835	–.080	–.126	–.027	–.166	–.013	16.540

The subjective judgment exercised in constructing the tests of the nine reading skills is reflected in the surprising extent to which several of the tests appear to be moderately "pure" factor measures. A word of caution must, however, be injected. Because some of the skills were judged to be more important than others in the reading process and because practical considerations governed to some extent the number of items used to measure each of the nine reading skills, the standard deviations of the initial variables differed considerably. And, since the initial matrix of variances and covariances used for the analysis reflected those differences, the coefficients in Table 5 must be interpreted with due regard for the magnitudes of the standard deviations of the nine initial skills. Scores in skill 1, for example, have a large standard deviation in comparison with the standard deviations of scores in the other skills. So a small component loading in skill 1 may be found to have more weight in a regression equation for obtaining scores in any one of the components than would be expected from an inspection of Table 5 alone.*

* Readers who are most familiar with the centroid method of factorial analysis have sometimes questioned this statement. A principal-axes analysis makes it possible to obtain very readily a given individual's score in any one of the components for which regression coefficients (or factor loadings) have been determined. For example, individual scores in component I may be obtained from the following regression equation:

$$C_I = .813(X_1) + .184(X_2) + .057(X_3) + .027(X_4) + .107(X_5) + .341(X_6) + .336(X_7) + .078(X_8) + .233(X_9).$$

In this equation, variables 6 and 7 have nearly identical regression coefficients, but we know that the standard deviation of variable 6 is 5.67 while that of variable 7 is 5.81. Therefore, variable 7 will have a slightly greater weight in determining an individual's score in component I than will variable 6 despite the

A study of the values in Table 5 (making due allowance for the magnitudes of the standard deviations of the initial variables) reveals that the nine components are rather readily identifiable in terms of the original nine reading skills. Component I is clearly word knowledge (skill 1). Its positive loadings in each of the nine basic reading skills reflect the fact that to read at all it is necessary to recognize words and to recall their meanings.

It is clear that word knowledge plays a very important part in reading comprehension and that any program of remedial teaching designed to improve the ability of students to understand what they read must include provision for vocabulary building. When one combines the evidence that word knowledge is so important an element in reading with the fact that the development of an individual's vocabulary is in large measure dependent on his interests and his background of experience, the relatively low correlations between reading tests in different subject-matter fields are understandable.* There is, however, no necessity to conclude that all of the fundamental factors of comprehension in reading are *not* involved in reading materials in various subject-matter fields.

Component II has been termed a measure of reasoning in reading. It has its highest positive loadings in the two reading skills that demand ability to infer meanings and to weave together several statements. It may seem puzzling at first that this component should have a strong negative loading in skill 1 (word knowledge), but consideration of the psychological meaning of components I and II indicates that this should be expected. The explanation undoubtedly lies in the fact that individuals who know accurately the meanings of a great many words are thereby given a head start toward getting the meaning of what they read. Therefore, if we are to measure reasoning in reading independently of word knowledge, we must give individuals who are deficient in word knowledge a "handicap" and then see how well they reason when they are placed on equal terms with their fellows in word knowledge. Component II apparently measures the ability to see the relationships of ideas.

* For data on this point see (5).

fact that the factor loadings of variables 6 and 7 in component I are almost the same.

A simple and convenient aid in interpreting the regression coefficients with proper regard for the sizes of the standard deviations of the initial variables is to construct a table containing each regression coefficient multiplied by the appropriate standard deviation of an initial variable. For example, the factor loading of skill 1 in component I (.813) would be multiplied by the standard deviation of skill 1 (11.61), yielding 9.4; the factor loading of skill 2 in component I (.184) would be multiplied by the standard deviation of skill 2 (3.25), yielding .6; and so on.

Component III is not so readily interpretable as most of the others, but it is clear that individuals who obtain high scores in this component focus their attention on a writer's explicit statements almost to the exclusion of their implications. Component IV measures chiefly the ability to identify a writer's intent, purpose, or point of view (skill 9). Individuals who obtain high scores in this component are less concerned with *what* a writer says than with *why* he says it. Such individuals should presumably be better able to detect bias and propaganda than individuals who obtain low scores in this component. Component V is composed principally of ability to figure out from the context the meaning of an unfamiliar word or to determine which one of several known meanings of a word is most appropriate in its particular contextual setting (skill 2). It is reasonable that it should be essentially unrelated to skill 1, which measures memory for isolated word meanings. The slight negative loadings of skills 6 and 7 in component V may result from the fact that the latter measures deductive reasoning, while skills 6 and 7 measure inductive processes.

Judging by its very high loading in skill 5, component VI seems to be largely a measure of ability to grasp the detailed statements in a passage. It is probably a fairly direct measure of the ability to get what I. A. Richards has called "the literal sense meaning" of a passage. Skill 5 was originally intended to measure this ability and the results of the analysis suggest that this ability is more than a name; it appears to be a real psychological entity distinct from other mental skills involved in reading. Component VII seems to be a measure principally of skill 3 (ability to follow the organization of a passage and to identify antecedents and references in it). The variance of this component consists of about 77% of the original variance of skill 3.

Component VIII measures specific knowledge of literary devices and techniques, and probably reflects the influence of training in English more than the other components do. Component IX is composed largely of ability to select the main thought of a passage; it may be considered a measure of ability in the synthesis of meaning. The variance of component IX comprises approximately 82% of the original variance of skill 4 (ability to select the main thought of a passage). Students who make high scores in component IX are presumably those who would be most capable of writing adequate summaries and précis of what they read.

Of the nine components described, all except components II, III, and IV can, for practical purposes, probably be measured satisfactorily by means of raw scores in one of the nine basic reading skills selected initially. Components V through IX account for only a small fraction of the total variance, but their variances are significantly different.* A number of the skills considered most important by authorities in the field of reading include independent elements that should be taught separately. It is not enough to assign learning exercises in reading that consist of passages followed by factual questions to be answered. Such exercises will not necessarily call the student's attention to the separate and essentially unrelated reading skills that he ought to master or give him sufficient practice in each one of them.

TABLE 6
Variance Ratios of Successive Components

Component	Degrees of Freedom	Variance	F
I	406	192.270	
			8.280
II	399	22.824	
			2.663
III	403	8.657	
			1.622
IV	399	5.282	
			1.387
V	401	3.828	
			1.158
VI	401	3.306	
			1.428
VII	403	2.327	
			1.181
VIII	400	1.956	
			1.944
IX	400	1.006	

* The writer is indebted to Professor T. L. Kelley for the development of a precise test of the variance ratios of components obtained by his iterative process. This test is described in the article by Professor Kelley that immediately follows.

The differences between the variances of successive components are all significant at the one-per-cent level with the exception of the differences between the variances of components V and VI, and VII and VIII; those differences are significant approximately at the five-per-cent level.

It should be noted that the variance-ratio test of the significance of the difference between component variances is permitted by the Kelley method but is not permitted by other methods of factorial analysis that are frequently employed.

Whether the variance of component IX (the smallest component) is significantly greater than would be yielded by chance may be determined by noting whether the reliability coefficient of component IX is significantly greater than zero. This is not established by the data. It is highly likely, however, that the variance of the next largest component is significantly greater than would be yielded by chance.

TABLE 7

Reliability Coefficients, Means, and Standard Deviations of the Six Independent Components Having Reliability Coefficients Substantially Greater Than Zero

Component	r_{11}	Mean	Standard Deviation
I	.94	46.30	13.87
II	.48	24.14	4.78
III	.28	.81	2.94
IV	.17	- .62	2.30
VII	.33	.27	1.53
VIII	.29	.70	1.40

Because individual scores in each of the independent components defined above can readily be estimated by using appropriate regression equations (Cf. ante, footnote following Table 5), the reliability coefficients of scores in the nine components have been determined empirically, using the same sample of 100 cases for which odd and even scores in each variable were obtained in computing the reliability coefficients of the nine initial variables.

Inspection of Table 7 reveals that only components I and II are measured with sufficient reliability to warrant their use for practical purposes. However, when the significance of the reliability coefficient of each one of the nine components is tested,* it becomes evident that useful measures of at least three additional components could certainly be provided by constructing the required number of additional items of the appropriate types. Since several of the components may be satisfactorily measured, for practical purposes, by raw scores in appropriate types of test items, construction of a large number of the indicated types of items has already been started. It is believed that these may be useful for instructional as well as for measurement purposes when they are employed in combination with other workbook materials.

Since useful measures of components I and II are already available, a profile chart for making a graphic record of scores in these two components has been prepared and is described in considerable detail elsewhere (9).

The correlations of components I and II with the Q and L scores derived from the *American Council on Education Psychological Ex-*

* The standard error of a split-half reliability coefficient, corrected by the Spearman-Brown formula, may be obtained by using Shen's formula,

$$\sigma_{r_{11}} = \frac{2(1 - r_{11})}{\sqrt{N}}.$$

amination and with the total score on the *Nelson-Denny Reading Test* have also been reported in the literature (9, 370-371). It is hoped that the relationships between components I and II and other well-known reading tests can be obtained, for if components I and II are regarded as fundamental abilities in reading it is of paramount importance to determine the extent to which the reading tests now commonly used in high schools and colleges actually measure each of these abilities.

The study reported here has explored one means of investigating the psychological nature of reading ability. It has suggested a means of determining the validity of tests of comprehension in reading. The results indicate that there is need for reliable tests to measure several of the nine basic skills that have been defined and for workbooks to aid in improving students' abilities in them. The need for correlating scores in existing reading tests with scores in several of the principal components seems obvious. And, not least, the study provides more detailed information regarding the skills measured by the *Cooperative Reading Comprehension Tests* than has heretofore been provided regarding the skills actually measured by any other widely used reading test.*

Finally, it is hoped that the data presented will draw attention to the importance of the mental skills involved in reading and act as a stimulus to further research in the fundamental factors of comprehension.

References

1. Adler, M. J. How to read a book. New York: Simon & Schuster, 1940.
2. Alderman, G. H. Improving comprehension in reading. *J. educ. Res.*, 1926, 13, 11-21.
3. Berry, B. T. Improving freshman reading habits. *Engl. J.*, College Edition, 1931, 20, 824-828.
4. Carroll, R. P. An experimental study of comprehension in reading. New York: Teachers College, Columbia University, 1927.
5. The Cooperative General Achievement Tests (Revised Series): Information concerning their construction, interpretation, and use. New York: Cooperative Test Service, 1940.
6. The Cooperative Reading Comprehension Tests: Information concerning their construction, interpretation, and use. New York: Cooperative Test Service, 1940.
7. The Cooperative Reading Comprehension Tests, Lower and Higher Levels; Forms Q, R, S, and T. New York: Cooperative Test Service.

* Frederick B. Davis, et al., *The Cooperative Reading Comprehension Tests, Lower and Higher Levels, Forms Q, R, S,* and *T*. Eight separate 40-minute reading tests are now available and are distributed by the Cooperative Test Service, 15 Amsterdam Avenue, New York, N. Y., a nonprofit agency of the American Council on Education.

8. Davis, Frederick B. Fundamental factors of comprehension in reading. Unpublished doctor's thesis on file at Harvard University Library, Cambridge, Massachusetts, 1941.

9. Davis, Frederick B. Two new measures of reading ability. *J. educ. Psychol.*, 1942, 33, 365-372.

10. Dewey, J. C. The acquisition of facts as a measure of reading comprehension. *Elem. Sch. J.*, 1935, 35, 346-348.

11. Feder, D. D. Comprehension maturity tests—a new technique in mental measurement. *J. educ. Psychol.*, 1938, 29, 597-606.

12. Gans, R. A study of critical reading comprehension in the intermediate grades. New York: Teachers College, Columbia University, 1940.

13. Gates, A. I. Methods of constructing and validating the Gates Reading Tests. *Teach. Coll. Rec.*, 1927, 29, 148-159.

14. Gates, A. I. The improvement of reading. New York: Macmillan Company, 1935. Ch. III and VII.

15. Gates, A. I. and Van Alstyne, D. General and specific effects of training in reading with observations on the experimental technique. *Teach. Coll. Rec.*, 1924, 25, 98-123.

16. Gray, W. S. Principles of method in teaching reading as derived from scientific investigation. National Society for the Study of Education, *Yearbook 18, Part II.* Bloomington, Illinois: Public School Publishing Company, 1919, 26-51.

17. Gray, W. S. The nature and types of reading. National Society for the Study of Education, *Yearbook 26, Part I.* Bloomington, Illinois: Public School Publishing Co., 1937, 23-40.

18. Gray, W. S. and Leary, B. E. What makes a book readable. Chicago: Univ. Chicago Press, 1935.

19. Hildreth, G. H. Learning the three R's. Minneapolis: Educational Publishers, Inc., 1936. Ch. III.

20. Hilliard, G. H. Probable types of difficulties underlying low scores in comprehension tests. *Univ. Iowa Studies in Education, II, No. 6.* Princeton, N. J.: Psychological Review Co., 1924.

21. Irion, T. W. H. Comprehension difficulties of ninth-grade students in the study of literature. New York: Teachers College, Columbia University, 1925.

22. Kelley, T. L. Essential traits of mental life. Cambridge, Massachusetts: Harvard Univ. Press. 1935.

23. Langsam, R. S. A factorial analysis of reading ability. *J. exp. Educ.*, 1941, 10, 57-63.

24. McCallister, J. M. Determining the types of reading in studying content subjects. *Sch. Rev.*, 1932, 40, 115-123.

25. Murphy, P. G. The role of the concept in reading ability. *Univ. Iowa Studies in Psychology*, XVII. Princeton, N. J.: Psychological Review Co., 1933, 21-73.

26. Pressey, L. W. and Pressey, S. L. A critical study of the concept of silent reading. *J. educ. Psychol.*, 1921, 12, 25-31.

27. Richards, I. A. Practical criticism. New York: Harcourt Brace, 1929.

28. Richards, I. A. How to read a page. New York: Harcourt Brace, 1942.

29. Richards, I. A. Interpretation in teaching. New York: Harcourt Brace, 1938.

30. Sangren, P. V. The improvement of reading through the use of tests. Lansing, Mich.: De Kleine, 1931.

31. Shank, S. Student responses in the measurement of reading comprehension. Cincinnati: C. A. Gregory, 1929.

32. Strang, R. Problems in the improvement of reading in high school and college. Lancaster, Pa.: Science Press Publishing Co., 1938. Chapter II.

33. Thorndike, E. L. The psychology of thinking in the case of reading. *Psychol. Rev.*, 1917, **24**, 220-234.

34. Thorndike, E. L. Reading as reasoning: A study of mistakes in paragraph reading. *J. educ. Psychol.*, 1917, **8**, 323-332.

35. Thorndike, E. L. The understanding of sentences. *Elem. Sch. J.*, 1917, **18**, 98-114.

36. Touton, F. C. and Berry, B. T. Reading comprehension at the junior college level. *Calif. Quarterly sec. Educ.*, 1931, **6**, 245-251.

37. Tyler, R. W. Measuring the ability to infer. *Educ. Res. Bull.*, 1930, **9**, 475-480.

38. Woody, C. Measurement of a new phase of reading. *J. educ. Res.*, 1923, **8**, 315-316.

39. Wyman, J. B. and Wendle, M. What is reading ability? *J. educ. Psychol.*, 1921, **12**, 518-531.

40. Yoakum, G. A. Reading and study. New York: Macmillan Co., 1928. Ch. II.

41. Zahner, L. C. The testing of comprehension. *Educ. Rec.* Supplement No. 13, 1940, 21, 71-89.

42. Zahner, L. C. An approach to reading through meanings. Ch. IV in Reading in General Education. Washington, D. C.: American Council on Education, 1940.

PSYCHOMETRIKA— VOL. 9, NO. 3
SEPTEMBER, 1944

A VARIANCE-RATIO TEST OF THE UNIQUENESS OF PRINCIPAL-AXIS COMPONENTS AS THEY EXIST AT ANY STAGE OF THE KELLEY ITERATIVE PROCESS FOR THEIR DETERMINATION

TRUMAN L. KELLEY
HARVARD UNIVERSITY

The immediately preceding article by Dr. Frederick B. Davis provides illustration of the method here given. Let the initial variables be:

$$\begin{array}{lc} & x_1 \ , \ x_2 \ , \cdots , \ x_m \ , \\ \text{with the degrees of freedom in each:} & N-1 , N-1 , \cdots , N-1 . \end{array}$$

Make a first rotation between x_1 and x_2, obtaining y_1 and y_2. The equation $y_1 = x_1 \cos\theta + x_2 \sin\theta$ constitutes a linear restriction, and, in addition to this, $\sum y_1 = 0$. Similarly for y_2, so we now have variables $y_1 , y_2 , x_3 , \cdots , x_m$, having the following degrees of freedom: $N-2 , N-2 , N-1 , \cdots , N-1$. Also, at this point y_1 and y_2 are independent, so a variance-ratio test is appropriate. The variance ratio is

$$F_{N-2,N-2} = \frac{V_{y_1}/(N-2)}{V_{y_2}/(N-2)} .$$

If the next rotation is between y_1 and x_3, we would have

$$\begin{array}{lccccc} \text{variables:} & z_1 , & y_2 , & z_3 , & x_4 , \cdots , & x_m , \\ \text{d.o.f.:} & N-3 , & N-2 , & N-2 , & N-1 , \cdots , & N-1 . \end{array}$$

The variance ratio

$$F_{N-3,N-2} = \frac{V_{z_1}/(N-3)}{V_{z_3}/(N-2)}$$

provides a precise test for the difference between these two variances. This last rotation may have introduced a very slight covariance, $c_{z_1 y_2}$, but hardly such as to vitiate a variance-ratio test between V_{z_1} and V_{y_2}, for the correlation between z_1 and y_2 is very low, and between the variances still lower, the correlation between variances being closely equal to the square of the correlation between variables.

When the iterative process is continued until all covariance in excess of the order of covariance yielded by chance, considering the size of the sample, is eliminated, the resulting variables are, within the limits of chance, the final principal-axes components, C_1, C_2, $\cdots$, C_m, and a variance-ratio test between the variance of any component and any other component is available. It is

$$F_{N-1-g,N-1-h} = \frac{V_{c_a}/(N-1-g)}{V_{c_b}/(N-1-h)},$$

in which g is the number of rotations involved in reaching the variable that is Component a, and h the number connected with the variable that is Component b.

PSYCHOMETRIKA— VOL. 9, NO. 3
SEPTEMBER, 1944

CONTRIBUTIONS TO THE MATHEMATICAL THEORY OF HUMAN RELATIONS VIII: SIZE DISTRIBUTION OF CITIES

N. RASHEVSKY
THE UNIVERSITY OF CHICAGO

An attempt is made to connect the distribution function of the sizes of the cities with the distribution functions of some other characteristics of the individuals in the society. Several theoretical possibilities are discussed and different relations are derived. A possible connection with some observed relations is discussed.

In a previous paper (13) we have outlined a theory of size distribution of cities, which represented a generalization of our previous theory of urbanization (4, 10). If $N(n)\,dn$ denotes the total number of individuals that inhabit all cities, the population of which is between n and $n + dn$, then $N(n)$ is determined essentially by the function $f(n, N)$, which represents the per capita production of goods in the cities of population n, N being the total population. The factors considered in that theory are essentially economic ones.

We shall now consider a different approach to the same problem. Before we do that, however, we shall briefly discuss the possible relation of the previous theory to available data.

G. K. Zipf (14) has found an interesting relation between the sizes of cities and their rank according to size. If we rank all cities consecutively in order of decreasing size and denote the population of a city of rank r by $n(r)$, then, according to Zipf, for many countries we have the relation

$$n(r) = \frac{C}{r}, \tag{1}$$

where C is a constant. By considering data for the United States and for Canada at different times, Zipf finds that relation (1) did not hold in the past, but has been gradually approached. He generalizes this by postulating that relation (1) is characteristic of a stable society, and is reached as a society passes from less stable to more stable configurations. In our opinion, such a postulate seems to lack any either empirical or theoretical evidence. It would therefore be of great interest if such a postulate should be contained as a deduction from a rational theory. This would throw light on the mechanism underlying the simple relation (1).

Inasmuch as the distribution functions studied by us previously do not involve the rank-order of the cities, we shall first investigate how Zipf's results translate themselves into our notations.

Denoting by $r(n)$ the rank of a city having a population n, equation (1) may also be written

$$r(n) = \frac{C}{n}. \tag{2}$$

Let $\bar{N}(n)\delta n$ be the number of cities whose population lies between n and $n + \delta n$, δn being a small but finite quantity. Consider instead of (2) any arbitrary function $r(n)$ which must, however, be decreasing with increasing n. For very large values of $r(n)$, the number of cities $N(n)\delta n$ will be large even for small δn. Since $r(n)$ increases discontinuously and always in steps of one, therefore $\bar{N}(n)\delta n$ equals $-\delta r(n)$, which corresponds to δn, for a given n.

$$\bar{N}(n)\ \delta n = -\delta r(n), \tag{3}$$

or

$$\bar{N}(n) = -\frac{\delta r(n)}{\delta n}. \tag{4}$$

For very large r's, we can take very small δn, and in the limit shall then find

$$\bar{N}(n) = -\frac{dr}{dn}. \tag{5}$$

The notion of a continuous distribution function $\bar{N}(n)$ breaks down for very large values of n, for there are only a few very large cities in each country. Nevertheless, far away from the tail end of the curve, the function $\bar{N}(n)$ may be practically determined. The rank-order notation has the advantage of covering the whole range of sizes.

The function $\bar{N}(n)$ is connected with the function $N(n)$ by the relation

$$\bar{N}(n) = \frac{N(n)}{n}. \tag{6}$$

Hence, if $r(n) = C/n$, then

$$\bar{N}(n) = \frac{C}{n^2}; \quad N(n) = \frac{C}{n} = r(n). \tag{7}$$

Denote by N the total population,

$$N = \int_0^\infty n\bar{N}(n)\,dn = \int_0^\infty N(n)\,dn\,. \tag{8}$$

A distribution function $N(n)$ such as given by equation (7) could be readily obtained from our previous theory by putting in (13)

$$f(n,N) = AnN\,, \tag{9}$$

where A is a constant. That would give us, according to equation (17) of (13),

$$N(n) = \frac{p}{An}, \tag{10}$$

p being determined as before from relation (8) of this paper.

Equation (9) means that the per capita production of goods is proportional to both n and N, a relation that does not seem to be very plausible. Because of relation (6) this would imply that the total production is proportional to $n^3\bar{N}^2$.

The simple theory developed before does not seem, therefore, to account for the relation found by Zipf. Inasmuch as not all countries follow Zipf's relation, and inasmuch as the above-mentioned theory may be modified and generalized, we should not discard it yet altogether. In a complex problem like this in which many factors probably enter, it may be advisable to discuss in abstracto different *conceivable* theoretical cases, without first worrying about actual data. A thorough classification of the theoretical possibilities may later on prove a help in deciding which of those possibilities or their combinations can actually be applied to observations. It is in this spirit of an abstract theoretical study that we shall discuss an alternative approach to the problem without prejudice to other possible approaches.

It is natural to attempt to connect the formation of cities with the presence of active groups in a society. A city may originate as an administrative center, in which case its formation will be closely related with the activity of an administrative active group, which we shall denote group *I,* as in previous papers (3-10). A city may also originate as a trade or industrial center, in which case it will be associated with another active group, which we previously denoted as group *II.* The stronger the administrative group *I,* the larger we may expect the principal city to be. On the contrary, a strong industrial group will result in the formation of large industrial cities. It is therefore natural to inquire whether the distribution function of city sizes may not be connected with the distribution function of the gradation of different types of activities within the population. In our earlier papers we began by considering continuous distribution functions for such quantities as coefficients of influence, etc., (1, 2). Subsequently, for mathematical convenience and as a practical approximation, we considered discontinuous distributions, assuming the population to be divided into completely active and completely passive ones without gradual transition. While such an assumption certainly does not correspond to reality, it serves as a good approximation. In some cases the results obtained by considering continuous distribution are found to be essentially the same as for the approximate discontinuous case (9).

Since this is a purely theoretical study, we shall not specify here what particular characteristics we do consider. We shall simply denote that characteristic by x and consider that in a population of N individuals the characteristic x is distributed according to some function $N(x)$. In whatever physical or psychophysical units we measure that characteristic x, we shall choose our units so that the maximum value of x in a given group is equal to 1. We then have

$$N = \int_0^1 \mathbf{N}(x)\,dx. \tag{11}$$

We have seen previously that such a group may break up into several smaller groups, if each individual associates only with those individuals whose x is not too remote from his own. Equations for determination of the size of those classes have been given previously (1, 2). We thus find that the whole population is divided into $n + 1$ groups, whose x's lie between $1 - x_1$, $x_1 - x_2$, $x_2 - x_3$, $\cdots x_n - 0$, with $x_k > x_{k+1}$. The total amount of x in a group r is given by

$$X(r) = \int_{x_r}^{x_{r-1}} x\mathbf{N}(x)\,dx\,. \tag{12}$$

The corresponding "populations" are given by

$$N(r) = \int_{x_r}^{x_{r-1}} \mathbf{N}(x)\,dx\,. \tag{13}$$

Suppose now that these groups will be segregated spatially. They will thus form separate communities of different sizes.

Let x denote any special ability such as executive, business, literary, etc., and consider the case where $\mathbf{N}(x)$ decreases with increasing x. The functions $N(r)$ and $X(r)$ may either decrease or increase with r. It is clear that we cannot set $n(r)$ proportional to $N(r)$ if the latter increases, for the largest community will be composed of individuals with the smallest x. If, however, $N(r)$ and $X(r)$ decrease with increasing r, we may consider the following situation.

The r-th group of $N(r)$ individuals may form a nucleus around which a number, $n'(r)$, of individuals of the lowest x gather to perform any activity directed by the $N(r)$ individuals, and thus form a community of $n(r) = N(r) + n'(r)$ individuals. We thus assume that the class of lowest x's, namely, that lying in the interval $x_n - 0$, is entirely passive. This class is also the most numerous one. The simplest assumption we may make about $n'(r)$ is that it is proportional to $X(r)$. We then have

$$n(r) = N(r) + \alpha X(r). \qquad (14)$$

If $N(r) << \alpha X(r)$, then we have approximately

$$n(r) = \alpha X(r). \qquad (15)$$

Thus $n(r)$ would be determined by $\mathbf{N}(x)$, and we may investigate what form of $\mathbf{N}(x)$ will give a prescribed $n(r)$, for example, the expression (1).

While such a case presents some theoretical interest, yet it can hardly be applied to actual cases. For it implies that the active group of each community has a different range of x's, the ranges for different cities *never overlapping.*

We may consider a somewhat more realistic assumption which is free from the above shortcoming. Let the group $N(1)$ form a nucleus around which there will be gathered $n'(1) = \alpha X(1)$ individuals, but let those $n'(1)$ individuals be taken from all the $N^{(1)} = N - N(1)$ individuals that are left outside of the group $N(1)$. Furthermore, let the contribution of individuals with a given x to $\alpha X(1)$ be proportional to the frequency with which those individuals occur. We have

$$N^{(1)} = \int_0^{x_1} \mathbf{N}(x)\,dx\,. \qquad (16)$$

The distribution function $N^{(2)}(x)$ of the individuals left after the $\alpha X(1)$ individuals have been subtracted from $N^{(1)}$ is, with the above assumption, equal to

$$N^{(2)}(x) = (1 - \frac{\alpha X(1)}{N^{(1)}})\,\mathbf{N}(x)\,. \qquad (17)$$

In other words, each class has lost a fraction $\alpha X(1)/N^{(1)}$ of individuals.

Now the group whose x lies between x_1 and x_2 has only

$$\overline{N}^{(2)} = \int_{x_2}^{x_1} N^{(2)}(x)\,dx \qquad (18)$$

individuals, and their total x is equal to

$$X'(2) = \int_{x_2}^{x_1} xN^{(2)}(x)\,dx\,. \tag{19}$$

This group will gather around it $\alpha X'(2)$ individuals, from the $N^{(2)} = N^{(1)} - \overline{N}^{(2)}$ individuals left outside of the group. We have

$$N^{(2)} = \int_{0}^{x_2} N^{(2)}(x)\,dx\,. \tag{20}$$

The distribution function of the remaining individuals is given by

$$N^{(3)}(x) = (1 - \frac{\alpha X'(2)}{N^{(2)}})\,N^{(2)}(x)\,, \tag{21}$$

and

$$X'(3) = \int_{x_3}^{x_2} xN^{(3)}(x)\,dx\,. \tag{22}$$

Thus we can consecutively calculate $N^{(1)}$, $\overline{N}^{(1)}$, $N^{(2)}$, $\overline{N}^{(2)}$, as well as $X(1)$, $X'(2)$, $X'(3)$, $\cdots$ etc., and thus find $n(r)$ from

$$n(r) = \overline{N}^{(r)} + \alpha X'(r). \tag{23}$$

This scheme may be generalized further by considering that as the distribution functions $N^{(i)}(x)$ change step by step, so will the set $x_1, x_2, \cdots, x_n$ change, according to equations developed previously (1). We shall have the interval $1 - x_1$ determined by $N^{(1)}(x) = \mathbf{N}(x)$. Instead of $x_1 - x_2$, we shall use in (18) an interval $x_1 - x'_2$, where x'_2 is determined by $N^{(2)}(x)$, and so forth.

The difficulty with actual calculation of such expressions lies in the circumstance that even the simplest forms of $\mathbf{N}(x)$ lead to transcendental equations for $x_1, x_2, \cdots, x_n$, which do not admit of closed solutions (1, 2). In order to get an idea as to how such expressions as (12), (13), (14), and (23) behave, we shall make a very crude approximation and consider all intervals $x_i - x_{i+1}$ as equal to a small constant Δ:

$$x_i - x_{i+1} = \Delta\,. \tag{24}$$

For very small values of Δ an approximate expression for $X(r)$ and $N(r)$ is easily obtained. We may put approximately

$$X(r) = \int_{1-r\Delta}^{1-(r-1)\Delta} x\mathbf{N}(x)\,dx = \Delta x\mathbf{N}(x), \tag{25}$$

and

$$N(r) = \int_{1-r\Delta}^{1-(r-1)\Delta} \mathbf{N}(x)\,dx = \Delta\mathbf{N}(x). \tag{26}$$

Remembering that for the r-th group x is approximately equal to $1 - r\Delta$, we find

$$X(r) = \Delta(1 - r\Delta)\mathbf{N}(1 - r\Delta), \quad N(r) = \Delta\mathbf{N}(1 - r\Delta). \tag{27}$$

We shall consider here, as a theoretically interesting case, an approximate distribution function

$$\mathbf{N}(x) = Ax^{-\nu}, \qquad \nu > 0, \tag{28}$$

which is suggested by Pareto's law. The relation (28) cannot hold physically for $x = 0$. Most likely $\mathbf{N}(0) = 0$, but it may also be that $\mathbf{N}(0) =$ Const. Except for exceedingly small values of x, the approximation may be very good. The exact expression for $\mathbf{N}(x)$ should satisfy relation (11), which also determines the constant A.

For very small values of Δ we have from (27)

$$X(r) = A\Delta(1 - r\Delta)^{1-\nu}; \qquad N(r) = A\Delta(1 - r\Delta)^{-\nu}. \tag{29}$$

If $X(r)$ is always to decrease with increasing r, we must have

$$0 < \nu < 1. \tag{30}$$

In the following we shall always consider that the restriction (30) is satisfied.

The exact expressions for $X(r)$ and $N(r)$ are obtained from (12) and (13):

$$\begin{aligned} X(r) &= \frac{A}{2-\nu}\{[1 - (r-1)\Delta]^{2-\nu} - [1 - r\Delta]^{2-\nu}\}; \\ N(r) &= \frac{A}{1-\nu}\{[1 - (r-1)\Delta]^{1-\nu} - [1 - r\Delta]^{1-\nu}\}. \end{aligned} \tag{31}$$

These expressions are to be used in either (14) or (15).

Since, however, $N(r)$ in this case increases with (r), equation (14) would have no meaning.

We now shall derive an explicit form for expression (23) based on (24) and (28), and show that the difficulty vanishes in this case. We have

$$\bar{N}^{(1)} = A\int_{1-\Delta}^{1} x^{-\nu}\,dx = \frac{A}{1-\nu}[1 - (1-\Delta)^{1-\nu}]; \tag{32}$$

$$N^{(1)} = A\int_{0}^{1-\Delta} x^{-\nu}\,dx = \frac{A}{1-\nu}[1 - (1-\Delta)^{1-\nu}]; \tag{33}$$

$$X'(1) = X(1) = A\int_{1-\Delta}^{1} x^{1-\nu}\,dx = \frac{A}{2-\nu}[1 - (1-\Delta)^{2-\nu}]. \tag{34}$$

Hence

$$N^{(2)}(x) = (1 - \frac{\alpha X(1)}{N^{(1)}})\, \mathbf{N}(x) = \\ (1 - \alpha \frac{[1 - (1 - \Delta)^{2-\nu}](1 - \nu)}{(1 - \Delta)^{1-\nu}(2 - \nu)})\, Ax^{-\nu}; \tag{35}$$

$$X'(2) = \int_{1-2\Delta}^{1-\Delta} xN^{(2)}(x)\,dx = \\ (1 - \alpha \frac{[1 - (1 - \Delta)^{2-\nu}](1 - \nu)}{(1 - \Delta)^{1-\nu}\,(2 - \nu)})\, \frac{A}{2\nu}\,[(1 - \Delta)^{2-\nu} - (1 - 2\Delta)^{2-\nu}]. \tag{36}$$

Define

$$a = \alpha \frac{1 - \nu}{2 - \nu}; \tag{37}$$

and

$$f(r) = (1 - a \frac{1 - (1 - \Delta)^{2-\nu}}{(1 - \Delta)^{1-\nu}})(1 - a \frac{(1 - \Delta)^{2-\nu} - (1 - 2\Delta)^{2-\nu}}{(1 - 2\Delta)^{(1-\nu)}}) \cdots \\ \cdots (1 - a \frac{[1 - (r - 1)\Delta]^{2-\nu} - [1 - r\Delta]^{2-\nu}}{(1 - r\Delta)^{1-\nu}}). \tag{38}$$

We shall now prove that in general

$$N^{(r)}(x) = Af(r - 1)x^{-\nu}; \tag{39}$$

$$X'(r) = \frac{A}{2 - \nu} f(r - 1)\{[1 - (r - 1)\Delta]^{2-\nu} - [1 - r\Delta]^{2-\nu}\}. \tag{40}$$

Expressions (39) and (40) hold for $r = 1$ and $r = 2$, as we have seen. We shall prove that if they hold for r, they hold for $r + 1$.

From (39) we have

$$N^{(r)} = \int_0^{1-r\Delta} N^{(r)}(x)\,dx = \frac{A}{1 - \nu} f(r - 1)\,(1 - r\Delta)^{1-\nu}, \tag{41}$$

$$N^{(r+1)}(x) = (1 - \frac{\alpha X(r)}{N^{(r)}})\, N^{(r)}(x)\,dx = \\ (1 - \alpha \frac{1 - \nu}{2 - \nu} \frac{[1 - (r - 1)\Delta]^{2-\nu} - [1 - r\Delta]^{2-\nu}}{(1 - r\Delta)^{1-\nu}}) f(r - 1)\,Ax^{-\nu}. \tag{42}$$

Because of (37) and (38), equation (42) may be written

$$N^{(r+1)}(x) = Af(r)x^{-\nu}. \tag{43}$$

We have further,

$$X'(r+1) = \int_{1-(r+1)\Delta}^{1-r\Delta} xN^{(r+1)}(x)\,dx\,,$$

which, because of (43), may be written

$$X'(r+1) = \frac{A}{2-\nu} f(r)\{[1-r\Delta]^{2-\nu} - [1-(r+1)\Delta]^{2-\nu}\}. \tag{44}$$

This proves (39) and (40).

We have

$$\bar{N}^{(r)} = \int_{1-r\Delta}^{1-(r-1)\Delta} N^{(r)}(x)\,dx = \frac{A}{1-\nu} f(r-1)\{[1-(r-1)\Delta]^{1-\nu} - [1-r\Delta]^{1-\nu}\}\,. \tag{45}$$

Hence, introducing (40) and (45) into (23):

$$n(r) = \frac{A}{1-\nu} f(r-1)\{[1-(r-1)\Delta]^{1-\nu} - [1-r\Delta]^{1-\nu}\} + \frac{\alpha A}{2-\nu} f(r-1)\{[1-(r-1)\Delta]^{2-\nu} - [1-r\Delta]^{2-\nu}\}. \tag{46}$$

The expression $f(n)$ simplifies considerably for very small values of Δ. Putting for any k,

$$k\Delta = y\,, \tag{47}$$

we have

$$[1-(k-1)\Delta]^{2-\nu} - [1-k\Delta]^{2-\nu} = [1-(y-\Delta)]^{2-\nu} - [1-y]^{2-\nu}. \tag{48}$$

For very small values of Δ, this is equal to

$$-\Delta\frac{d}{dy}(1-y)^{2-\nu} = (2-\nu)\Delta(1-y)^{1-\nu} = (2-\nu)\Delta(1-k\Delta)^{1-\nu}. \tag{49}$$

Hence, because of (48), (49), and (37),

$$1 - a\,\frac{[1-(k-1)\Delta]^{2-\nu} - [1-k\Delta[^{2-\nu}}{(1-k\Delta)^{1-\nu}} = 1 - \alpha(1-\nu)\Delta; \tag{50}$$

and, because of (38):

$$f(r) = [1-\alpha(1-\nu)\Delta]^{r}. \tag{51}$$

Introducing (51) into (40), and transforming the expression in braces of (40) according to (48) and (49), we have:

$$X'(r) = A\Delta(1 - r\Delta)^{1-\nu}[1 - \alpha(1 - \nu)\Delta]^{r-1}. \qquad (52)$$

Since physically we must have

$$0 << 1 - \alpha(1 - \nu)\Delta < 1, \qquad (53)$$

therefore, comparing (52) with (29), we see that $X'(r)$ decreases more rapidly with r than $X(r)$. This should be physically so, because while $X(r)$ refers to the group formed of all individuals who have an x between $1 - r\Delta$ and $1 - (r - 1)\Delta$, $X'(r)$ refers to the group formed by individuals left within that interval after subtracting the amount which contributed the $r - 1$ preceding n's. Hence $X(r) > X'(r)$.

It should be noticed that $X(r)$ decreases with r less rapidly than $1/r$, but $X'(r)$ decreases more rapidly than $1/r$.

By a similar procedure using (48), (49), and (51), we obtain from (45) for very small values of Δ:

$$\bar{N}^{(r)} = A\Delta(1 - r\Delta)^{-\nu}[1 - \alpha(1 - \nu)\Delta]^{r-1}. \qquad (54)$$

Introducing (52) and (54) into (23), we find:

$$n(r) = A\Delta[1 - \alpha(1 - \nu)\Delta]^{r-1}\{(1 - r\Delta)^{-\nu} + \alpha(1 - r\Delta)^{1-\nu}\}. \qquad (55)$$

The variation of $n(r)$ with r is rather complicated. For small values of r the term $[1 - \alpha(1 - \nu)\Delta]^{r-1}$ decreases more rapidly than $(1 - r\Delta)^{-\nu}$ increases. Therefore, for small values of r, $\bar{N}^{(r)}$ decreases, but less rapidly than $X'(r)$. The quantity $n(r)$ decreases also. Since, however, for $r = 1/\Delta$, the term $(1 - r\Delta)^{-\nu}$ becomes infinite, $n(r)$ has a minimum for some value of $r = r_m$.. Since by definition r is the rank-order of *decreasing* sizes, such a situation would be physically absurd. This difficulty may be avoided by the following consideration. Equation (55) is based on the approximation expressions (52) and (54), which cannot be applied for values of r that are close to $1/\Delta$. It must be remembered that r varies from 1 to $1/\Delta$ only. If the parameters in equation (55) can be chosen so that the value r_m for which $n(r)$ has a minimum is greater than $1/\Delta - 1$, then the above difficulty will be avoided. We shall now prove that this can be done.

Denote

$$\gamma = 1 - \alpha(1 - \nu)\Delta; \quad 0 < \gamma < 1. \qquad (56)$$

Equation (55) now becomes

$$n(r) = A\Delta\gamma^{r-1}\{(1 - r\Delta)^{-\nu} + \alpha(1 - r\Delta)^{1-\nu}\}. \qquad (57)$$

We have

$$\frac{dn(r)}{dr} = A\Delta\gamma^{r-1}(1-r\Delta)^{-\nu}[\nu\Delta(1-r\Delta)^{-1} + \log\gamma$$
$$- \alpha(1-\nu)\Delta + \alpha(1-r\Delta)\log\gamma] \,. \tag{58}$$

The value r_m is defined by $dn(r)/dr = 0$ or

$$\nu\Delta + (1-r\Delta)[\log\gamma - \alpha(1-\nu)\Delta]$$
$$+ \alpha(1-r\Delta)^2 \log\gamma = 0 \,. \tag{59}$$

Introduce the new variable

$$z = 1 - r\Delta; \quad z_m = 1 - r_m\Delta \,. \tag{60}$$

We then have

$$z_m{}^2\, \alpha \log(1/\gamma) + z_m[\log(1/\gamma) + \alpha(1-\nu)\Delta] - \nu\Delta = 0 \,. \tag{61}$$

Equation (61) gives

$$z_m = [(1/2)\log(1/\gamma)]\{-[\log(1/\gamma) + \alpha(1-\nu)\Delta]$$
$$+ \sqrt{[\log(1/\gamma) + \alpha(1-\nu)\Delta]^2 + 4\alpha\nu\Delta\log(1/\gamma)}\} \,. \tag{62}$$

The positive sign must be taken before the square root because $z_m > 0$.

If we wish to have $r_m < 1/\Delta - 1$, then we must have

$$z_m > \Delta \,. \tag{63}$$

Inequality (63) will be satisfied if γ is made sufficiently small. To show this, we make use of (56) and write (62) thus:

$$z_m = [(1/2)\log(1/\gamma)]\{-[\log(1/\gamma) + 1 - \gamma]$$
$$+ \sqrt{[\log(1/\gamma) + 1 - \gamma]^2 + \frac{4(1-\gamma)\nu}{1-\nu}\log(1/\gamma)} \,. \tag{64}$$

As γ becomes very small, $\log(1/\gamma)$ becomes very large. Thus we may neglect $1 - \gamma$ in the expression in brackets, as well as neglect γ as compared with 1. We then have:

$$z_m = [1/2)\log(1/\gamma)]\{-\log\frac{1}{\gamma}$$
$$+ \log(1/\gamma)\sqrt{1 + \frac{4\nu}{1-\nu}\,\frac{1}{\log(1/\gamma)}}\Bigg\} \,. \tag{65}$$

$1/\log(1/\gamma)$ being now a very small quantity, we may expand the expression under the square root sign, keeping only linear terms. We thus find

$$z_m = [(1/2)\log(1/\gamma)]\,\frac{2\nu}{1-\nu}. \tag{66}$$

By making γ sufficiently small, we can always satisfy inequality (63).

But a small γ means a sufficiently large α. Hence inequality (63) may be satisfied by taking a sufficiently large α, though not large enough to make γ negative.

Thus with a proper choice of α, $n(r)$ as given by (55) will be monotonically decreasing with r, within the range $1 \gtreqqless r \gtreqqless 1/\Delta - 1$. We may now investigate under what conditions, if any, $n(r)$ will vary within a wide range *approximately* as $1/r$, so as to satisfy relation (1).

The general theory of the breaking up of a social group into classes, as developed in previous papers (1, 2), is based on the assumption that only such individuals associate with each other for whom the difference $(x' - x)^2$ is less than a certain quantity $\Delta^2{}_0$. We had as a criterion,

$$(x' - x)^2 < \Delta^2{}_0. \tag{67}$$

We shall now consider a different criterion which is perhaps somewhat more realistic. We shall assume, namely, that it is not the difference $x' - x$, but the ratio x'/x, which determines whether or not two individuals associate with each other. The plausibility of such an assumption is suggested by the following considerations.

An individual with an income of $100,000 is likely to associate with another individual whose income is $75,000, but an individual with an income of $26,000 is not likely to associate with an individual having an income of $1,000. The difference is the same in both cases, but the ratios are different. Similarly, an executive or a politician who controls directly or indirectly 10,000 individuals will associate with another one who controls 6,000 individuals, but an executive having control over 5,000 individuals will not associate with a foreman having control over 25 individuals.

Instead of (67) we may now put

$$(\log x' - \log x)^2 < \Delta_0{}^2, \tag{68}$$

and instead of equation (13) of (2), we shall determine x from the equation

$$\int_x^1 \int_{x'}^1 [(\log x' - \log x)^2 - \Delta_0{}^2]\, N(x) N(x')\, dx dx'. \tag{69}$$

Equations of similar form will determine x_2, x_3, etc.

We run again into the same difficulty as before, namely, the equations determining x_i are transcendental. We shall therefore again use a very rough approximation corresponding to (24) of the previous case. We shall put, namely,

$$x_i = \beta^i, \quad 0 < \beta < 1. \tag{70}$$

For $\mathbf{N}(x)$ we shall again use (28).

We find now, with reference to (12) and (13),

$$X(r) = A \int_{\beta^r}^{\beta^{r-1}} x^{1-\nu}\, dx = \frac{A(1-\beta^{2-\nu})}{2-\nu} \beta^{(2-\nu)(n-1)}; \tag{71}$$

$$N(r) = A \int_{\beta^r}^{\beta^{r-1}} x^{-\nu}\, dx = \frac{A(1-\beta^{1-\nu})}{1-\nu} \beta^{(1-\nu)(n-1)}. \tag{72}$$

Now both $X(r)$ and $N(r)$ decrease monotonically with r, so that equation (14) can be used, giving

$$n(r) = A\Big(\frac{1-\beta^{1-\nu}}{1-\nu} \beta^{(1-\nu)(n-1)} + \alpha \frac{1-\beta^{2-\nu}}{2-\nu} \beta^{(2-\nu)(n-1)}\Big). \tag{73}$$

It is readily seen that $n(r)$ decreases much more rapidly than $1/r$.

We now calculate $X'(r)$, $N^{(r)}(x)$, and $\overline{N}^{(r)}$. We have

$$X'(1) = A \int_{\beta}^{1} x^{1-\nu}\, dx = \frac{A(1-\beta^{2-\nu})}{2-\nu}; \tag{74}$$

$$N^{(1)} = A \int_{0}^{\beta} x^{-\nu}\, dx = \frac{A\beta^{1-\nu}}{1-\nu}; \tag{75}$$

$$\begin{aligned} N^{(2)}(x) &= \Big(1 - \frac{\alpha X'(1)}{N^{(1)}}\Big) \mathbf{N}(x) \\ &= A\Big(1 - \alpha \frac{(1-\nu)(1-\beta^{2-\nu})}{(2-\nu)\beta^{1-\nu}}\Big) x^{-\nu}; \end{aligned} \tag{76}$$

$$\begin{aligned} X'(2) &= \int_{\beta^2}^{\beta} x N^{(2)}(x)\, dx \\ &= \Big(1 - \alpha \frac{(1-\nu)(1-\beta^{2-\nu})}{(2-\nu)\beta^{1-\nu}}\Big) \frac{A(1-\beta^{2-\nu})}{2-\nu} \beta^{2-\nu}. \end{aligned} \tag{77}$$

Define

$$b = \alpha \frac{(1-\nu)(1-\beta^{2-\nu})}{2-\nu}; \tag{78}$$

and

$$f_1(y) = (1 - b\beta^{\nu-1})(1 - b\beta^{\nu})(1 - b\beta^{\nu+1}) \cdots (1 - b\beta^{\nu+y}). \qquad (79)$$

We shall now prove that in general

$$N^{(r)}(x) = Af_1(r-3)x^{-\nu};$$

$$X'(r) = \frac{Ab}{\alpha(1-\nu)} f_1(r-3)\beta^{(2-\nu)(r-1)}. \qquad (80)$$

We have from (80):

$$N^{(r)} = \int_0^{\beta^r} N^{(r)}(x)\,dx = \frac{A}{1-\nu} f_1(r-3)\beta^{r(1-\nu)}. \qquad (81)$$

Hence

$$\begin{aligned} N^{(r+1)}(x) = (1 - \frac{\alpha X'(r)}{N^{(r)}})\, N^{(r)}(x) &= (1 - b\beta^{\nu+r-2})N^{(r)}(x) \\ &= Af_1(r-2)x^{-\nu}, \end{aligned} \qquad (82)$$

and

$$\begin{aligned} X'(r+1) = \int_{\beta^{r+1}}^{\beta^r} xN^{(r+1)}(x)\,dx &= f_1(r-2)\,\frac{A(1-\beta^{2-\nu})}{2-\nu}\beta^{r(2-\nu)} \\ &= \frac{Ab}{\alpha(1-\gamma)} f_1(r-2)\beta^{(2-\nu)r}. \end{aligned} \qquad (83)$$

Since (81) holds for $r = 2$, it therefore holds for any r. We also have

$$\bar{N}^{(r)} = \int_{\beta^r}^{\beta^{r-1}} N^{(r)}(x)\,dx = \frac{A(1-\beta^{1-\nu})}{1-\nu} f_1(r-3)\beta^{(r-1)(1-\nu)}. \qquad (84)$$

Introducing (80) and (81) into (23), we find

$$n(r) = \frac{A}{1-\nu} f_1(r-3)\beta^{(r-1)(1-\nu)}(1 - \beta^{1-\nu} + b\beta^{r-1}). \qquad (85)$$

$n(r)$ decreases monotonically with r, but more rapidly than $1/r$.

We may consider a more general assumption, namely that β varies from class to class. Denoting by $\beta_1, \beta_2, \cdots, \beta_i$ a sequence of numbers such that

$$0 < \beta_i < 1, \qquad (86)$$

we may put

$$x_i = \beta_1 \beta_2 \cdots \beta_i. \qquad (87)$$

In that case

$$X(r) = A \int_{\beta_1\beta_2\ldots\beta_r}^{\beta_1\beta_2\ldots\beta_{r-1}} x^{1-\nu}\, dx = \frac{A(1-\beta_r^{2-\nu})}{2-\nu} (\beta_1 \cdots \beta_r)^{2-\nu}. \qquad (88)$$

$X(r)$ decreases with r, and by a proper choice of the sequence $\beta_1\ \beta_2 \cdots \beta_i$ it may be made to decrease as $1/r$. Equation (15) would then satisfy (1). A similar assumption may be studied for the more general case involving $N^{(r)}$, $X'(r)$, and equation (23). This shall be done elsewhere.

The author is indebted to Professor Alston S. Householder for a critical discussion of this paper.

REFERENCES

1. Rashevsky, N. Outline of a mathematical theory of human relations. *Philosophy of Science*, 1935, **2**, 413-429.
2. Rashevsky, N. Further contributions to the mathematical theory of human relations. *Psychometrika*, 1936, **1**, 21-31.
3. Rashevsky, N. Studies in mathematical theory of human relations. *Psychometrika*, 1939, **4**, 221-239.
4. Rashevsky, N. Studies in mathematical theory of human relations, II. *Psychometrika*, 1939, **4** 283-299.
5. Rashevsky, N. Contributions to the mathematical theory of human relations. III. *Psychometrika*, 1940, **5**, 203-210.
6. Rashevsky, N. Contributions to the mathematical theory of human relations. IV. Outline of a mathematical theory of individual freedom. *Psychometrika*, 1940, **5**, 299-303.
7. Rashevsky, N. Note on the mathematical theory of interaction of social classes. *Psychometrika*, 1941, **6**, 43-47.
8. Rashevsky, N. On the variation of the structure of a social group with time. *Psychometrika*, 1941, **6**, 273-277.
9. Rashevsky, N. and A. S. Householder. On the mutual influence of individuals in a social group. *Psychometrika*, 1941, **6**, 317-321.
10. Rashevsky, N. Contributions to the mathematical theory of human relations. V. *Psychometrika*, 1942, **7**, 117-134.
11. Rashevsky, N. Further studies in the mathematical theory of interaction of individuals in a social group. *Psychometrika*, 1942, **7**, 225-232.
12. Rashevsky, N. Contributions to the mathematical theory of human relations. VI. Periodic fluctuations in the behavior of social groups. *Psychometrika*, 1943, **8**, 81-85.
13. Rashevsky, N. Contributions to the mathematical theory of human relations. VII. Outline of a mathematical theory of the size of cities. *Psychometrika*, 1943, **8**, 87-90.
14. Zipf, G. K. National unity and disunity. Bloomington, Ind.: The Principia Press. 1941.

PSYCHOMETRIKA— VOL. 9, NO. 3
SEPTEMBER, 1944

MOORE, UNDERHILL AND CALLAHAN, CHARLES C. *Law and Learning Theory: A Study in Legal Control.* New Haven: Yale Law Journal Co., Inc. Pp. 136. 1943.

A REVIEW

"The behavior selected for study was the familiar and everyday practice of parking automobiles in city streets. It was selected . . . because it is a form of behavior that occurs frequently, can be observed easily, be measured in terms of frequency and duration of occurrence, can be modified by posting signs indicating whether or not parking is legally allowed and if so for how long, and above all it can be observed without the slightest interference with it. By securing the cooperation of the local police department, the authors were able to compare parking behavior under three main conditions: (1) when parking was unrestricted, (2) restricted or prohibited but 'no tagging' for over-parking, (3) restricted and enforced by 'tagging.' "

Findings are interpreted in terms of "learning theory," which "conceives of the overt behavior of an individual, acting either alone or in a group of other individuals, as behavior which he has learned or is learning to perform. His behavior is determined by the relation between four factors—drive, cue, response, and reward—which relation he has learned, or is learning."

The authors show "that changes in the frequency and duration of parking behavior of an unselected sample of a population can be predicted by the usc of their empirical formulas. The question that remains unanswered is whether or not these empirical formulas can be *derived deductively* from any set of basic postulates."

STEUART HENDERSON BRITT
Washington, D. C.